



## AI for real-world network operation

### **WP1 – Design of a holistic framework for AI in critical network infrastructures**

#### D1.1 – AI4REALNET framework and use cases



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		D1.1 – AI4REALNET framework and use cases
TYPE	R — Document, report	
DISTRIBUTION LEVEL	PU - Public	
DUE DELIVERY DATE	30/09/2024	
DATE OF DELIVERY	30/09/2024	
VERSION	V1.0	
DELIVERABLE RESPONSIBLE	INESC TEC	
AUTHOR (S)	Ricardo Bessa (INESC TEC), Mouadh Yagoubi (IRTSX), Milad Leyliabadi (IRTSX)	
OFFICIAL REVIEWER/s	Herke van Hoof (UvA), Alberto Castagna (ENLITEAI), Anton Fuxjäger (ENLITEAI)	

## DOCUMENT HISTORY

VERSION	AUTHORS	DATE	CONTENT AND CHANGES
0.1	Ricardo Bessa	06/07/2024	Inclusion of the use cases contribution (Task 1.2)
0.2	Mouadh Yagoubi	20/07/2024	The first draft version of the conceptual framework (Task 1.1)
0.3	Ricardo Bessa	31/07/2024	Revision of the first version of the framework and comments
0.4	Milad Leyliabadi	11/09/2024	Addition and harmonization of all references, improvement of schemes and resolving the comments
0.5	Herke van Hoof	14/09/2024	Improvement of the knowledge-assisted AI part
0.6	Ricardo Bessa	14/09/2024	The first version was released for official review
0.7	Ricardo Bessa, Milad Leyliabadi, Manuel Schneider, Anna Fedorova, and all	27/09/2024	Corrections according to reviewer's comments
1.0	Ricardo Bessa	30/09/2024	Final version

## ACKNOWLEDGEMENTS

NAME	PARTNER
Milad Leyli-Abadi	IRTSX
Maroua Meddeb	IRTSX
Daniel Boos	SBB
Clark Borst	TU Delft
Alberto Castagna	ENLITEAI
Adrian Egli	SBB
Andrina Eisenegger	FHNW
Anton Fuxjäger	ENLITEAI
Samira Hamouche	FHNW
Mohamed Hassouna	FHG & UKASSEL
Bruno Lemetayer	RTE
Antoine Marot	RTE
Roman Liessner	DB
Manuel Schneider	FLATLAND
Irene Sturm	DB
Julia Usher	ZHAW
Herke Van Hoof	UvA
Jan Viebahn	TENNET
Sjoerd Kop	TENNET
Toni Waefler	FHNW
Joaquim Geraldes	NAV
Cristina Felix	NAV
Hélio Sales	NAV
Giulia Leto	TU Delft
Joost Ellerbroek	TU Delft
Ricardo Chavarriaga	ZHAW
Viola Schiaffonati	POLIMI
Giacomo Zanotti	POLIMI
Jonas Lundberg	LiU
Anna Fedorova	ZHAW

## 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

This document establishes the main foundations of the AI4REALNET project, in particular, the following key outcomes:

- The formal specification of domain-specific use cases (UCs), replicating real-world operating scenarios involving human operators to apply innovative AI-based methods. This is complemented by a comprehensive set of quantitative and qualitative key performance indicators (KPIs) addressing socio-technical aspects.
- The development of a domain-general conceptual framework that integrates social sciences and humanities (including psychology, ethics, and philosophy), human-centered design sciences, artificial intelligence (AI), and domain-specific expertise applied to critical infrastructures (power grid, railway network, and air traffic).

## USE CASES

The methodology for formally documenting the project’s UCs involved several key steps:

1. Identifying tasks at the network operators (i.e., RTE, TenneT, DB, SBB, and NAV) that are evolving or emerging with the development of AI and digital technologies.
2. Refining this list and draft descriptions through consortium meetings, workshops with external stakeholders, public webinars, and literature reviews related to AI’s impact on the three network infrastructure domains.
3. Basing the work on a thorough analysis of each network operator’s roadmap, internal organization, and the current regulatory framework, including anticipated short- and medium-term developments.

The network operators were responsible for describing the UCs and capturing the associated functional and non-functional requirements, supported by their domain experts and reviewed by R&D partners. The AI4REALNET project developed a template document based on the work presented in ISO/IEC TR 24030. Additionally, each UC identified a set of KPIs and specific business/task objectives aimed at capturing technical, economic, social, and human dimensions. The Assessment List for Trustworthy Artificial Intelligence (ALTAI) was also adopted as a comprehensive self-assessment tool across various dimensions. This tool was used to capture non-functional requirements related to trustworthy AI in the design of the UCs.

The following UCs were identified:

**UC1.Power Grid. AI assistant supporting human operators’ decision-making in managing power grid congestion:** Provide a human operator with remedial action recommendations aimed at safely managing overloads on the electrical lines and easing the workload of the human operator.

**UC2.Power Grid. Sim2Real, transfer AI-assistant from simulation to real-world operation:** Provide a human operator with remedial action recommendations, considering a transfer from training (digital) to real-world environments.

**UC1.Railway. Automated re-scheduling in railway operations:** The re-scheduling task is performed in a highly automated manner by an AI-based re-scheduling system. It observes the real-time state of all the trains and tracks in the control area of interest and automatically detects the need to intervene, decides on an intervention, and executes this intervention.

**UC2.Railway: AI-assisted human re-scheduling in railway operations:** Assist the human dispatcher in railway operations in re-scheduling train runs to fulfill all offered services and minimize delays for the customer.

**UC1.ATM. Airspace sectorization assistant:** Partially and fully automate the sectorization process to assist the supervisor in deciding when and how to split and merge sectors to balance the workload of Tactical Air Traffic Controllers (ATCOs).

**UC2.ATM: Flow & airspace management assistant:** Provide advice to ATCO about deviations with better sector capacity adherence and performance measured by an indicator of environmental area. Also, consider the need to review the sectorization plan due to the activation of military areas and required trajectory efficient deviations.

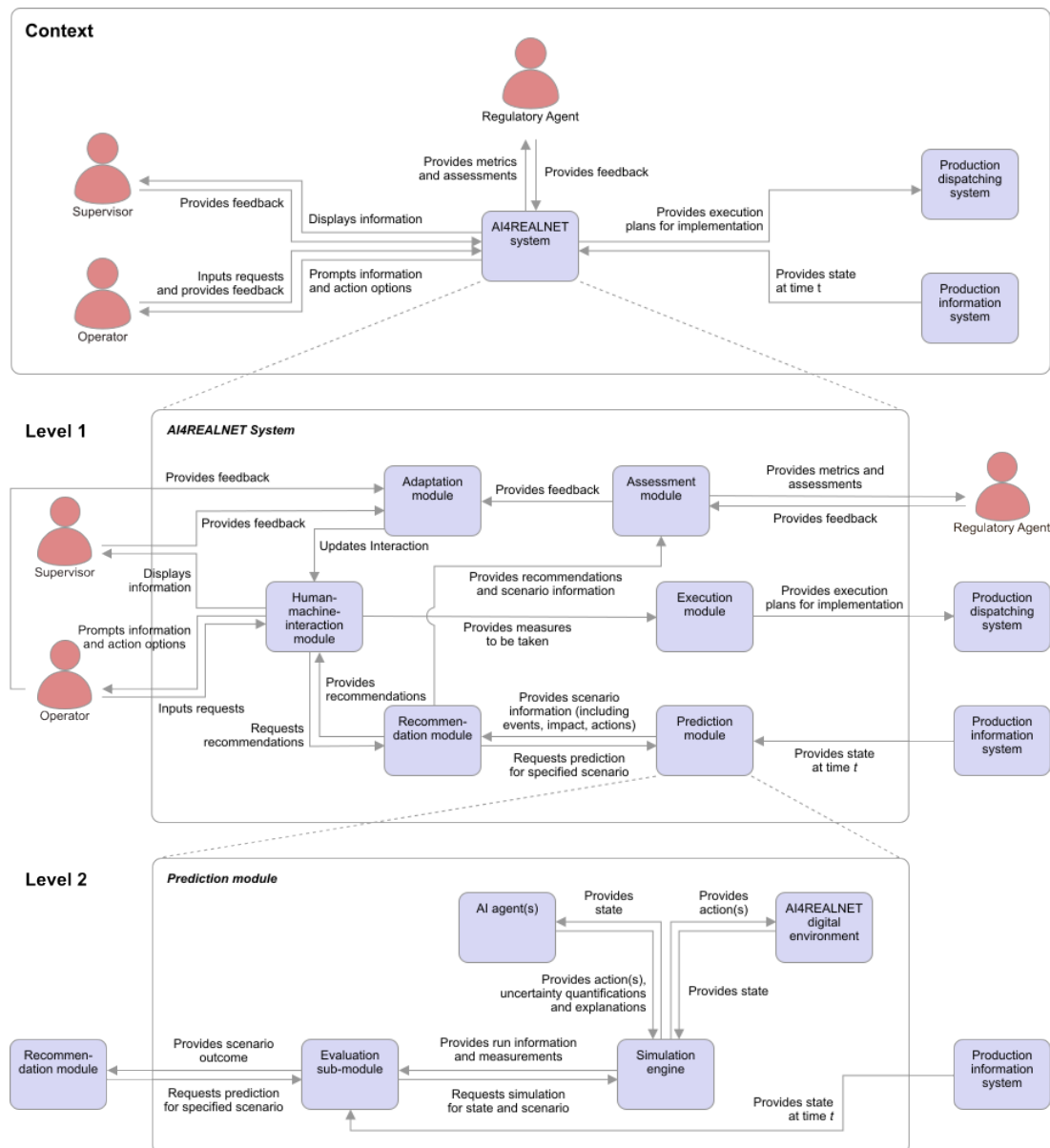
## CONCEPTUAL FRAMEWORK

The AI4REALNET consortium adopted an interdisciplinary approach to develop its conceptual framework, integrating traditionally distinct fields such as psychology and cognitive engineering. This enabled the study of expert collaborative decision-making in complex scenarios, where automation plays a role, and the development of effective design and evaluation criteria to support human decision-making. The framework also drew on mathematics, decision theory, computer science, and specialized engineering domains, particularly energy and mobility. Systems engineering and theories adapted for trustworthy AI integration were used in designing the system's operational, functional, and logical architecture to meet both functional and non-functional requirements of the UCs.

The conceptual framework is structured into various layers:

- The **context**, characteristics, impacts, and **decision environment** for critical network infrastructures are discussed based on the UC scenarios. This describes, in a unified way, the similarities and dissimilarities of the operating decision processes in the three critical infrastructures.
- **Decision-making** from a **socio-technical systems perspective (human agent)**, aiming for joint optimization to increase the whole system's performance. Namely, to take requirements derived from characteristics of the social sub-system (i.e., human factors) and be able to exploit AI capabilities and potentials, the social sub-system also needs to be designed accordingly.
- **Decision-making** process from the **AI perspective (AI Agent)** and the corresponding strategies and methods. It elaborates on the different characteristics an AI-based model should possess for efficient interactions between AI and human decision-makers in various situations and modes of interactions.
- **Epistemological and normative foundations of trustworthy AI** and analysis of the different components of risk and their application to AI, focusing on safety-critical systems.

At the **system design level for human-AI interaction**, the focus shifts to translating these layers into practical applications. To enhance the connection between research questions and real-world applications, we developed a high-level conceptual prototype – the AI4REALNET system – that allows us to test and refine ideas, ensuring research outcomes meet practical needs. It will evolve during the project and serve as initial design guidelines for future applications. This system offers a hierarchical representation of the system from a technical perspective. The figure below shows the scope, context, and high-level view of the AI4REALNET AI-based (conceptual) system.



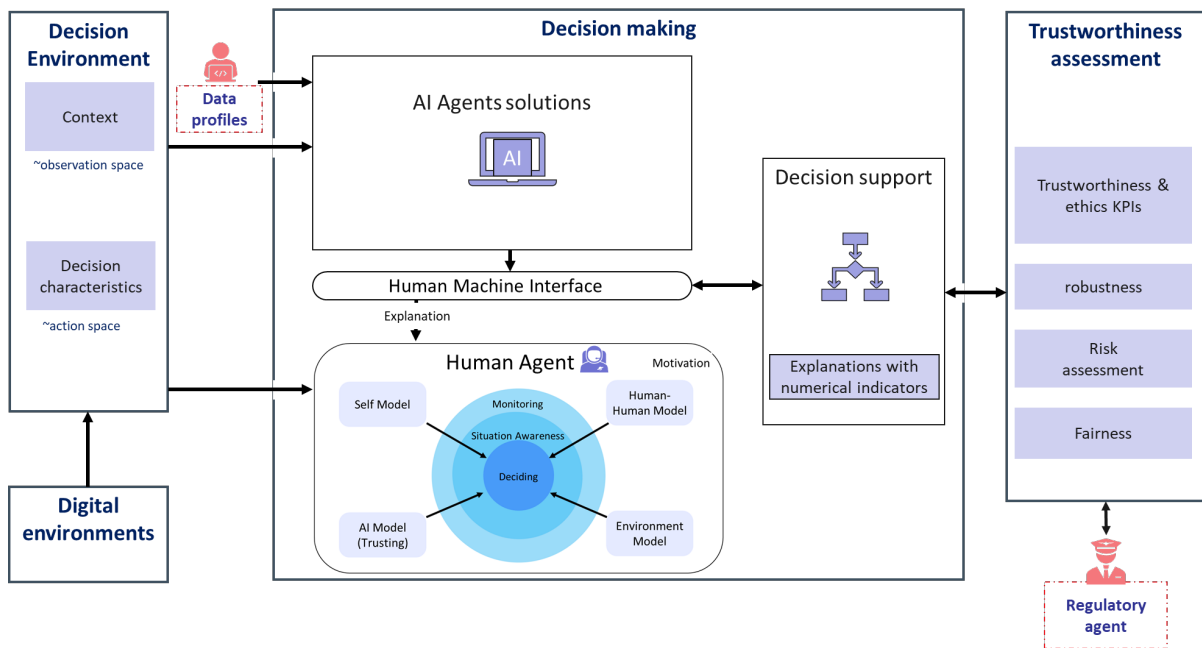
The system’s **context** includes neighboring systems to provide real-time operational information (production information system) and implement decisions taken within the system in live operations (production dispatching system). Further, users, such as operators, supervisors, and regulatory agents, are also part of the context and interact with the system.

In **Level 1**, the system is organized into modules based on function. The Human-Machine-Interaction module manages how AI interacts with humans, providing notifications, contextual information, and

assisting with tasks. The Adaptation module recognizes situations, adjusts human-AI interaction, and updates AI models based on feedback. The Prediction module forecasts events, assesses their impact, and stores important data. The Recommendation module suggests actions and explanations to operators, while the Execution module implements operational actions. Lastly, the Assessment module evaluates AI behavior, robustness, and fairness.

In **Level 2**, the Prediction module, central to the system, includes an evaluation sub-module, simulation engine, AI agents, and the AI4REALNET digital environment. It receives current system data and requests simulations to predict events and consequences. Simulation results are evaluated and sent to the recommendation module, with all relevant data stored for future use.

The generic process is illustrated with a high-level overview of the interactions between different sub-systems, each broken down into specific functions. The figure below presents the logical architecture of the conceptual framework for an AI assistant (with humans maintaining full control), which is one of the three scenarios considered in AI4REALNET: *AI-assistant to human (human in control)*, *joint human-AI decision-making (including human-AI co-learning)*, and *autonomous AI (human as a supervisor)*.



Finally, this design process addresses key aspects such as robustness, uncertainty quantification, knowledge-assisted AI, human-AI collaboration, explainability, and multi-objective reinforcement learning, creating a unified conceptual framework for different modes of human-AI interaction.

# TABLE OF CONTENTS

SUMMARY	4
TABLE OF CONTENTS	8
LIST OF FIGURES	10
LIST OF TABLES	12
ABBREVIATIONS AND ACRONYMS	13
1. INTRODUCTION	14
2. USE CASES AND METHODOLOGY	16
2.1 METHODOLOGY	16
2.2 AI4REALNET USE CASES	22
2.3 KEY PERFORMANCE INDICATORS	36
2.4 ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE (ALTAI)	45
3. CONCEPTUAL FRAMEWORK	57
3.1 CONTEXT AND DECISION ENVIRONMENT	58
3.2 DECISION-MAKING PROCESS	61
3.3 EPISTEMOLOGICAL AND PHILOSOPHICAL FOUNDATIONS OF TRUSTWORTHY AI	120
4. CONCLUDING REMARKS	129
REFERENCES	131
ANNEX 1 – USE CASE TEMPLATE	140
ANNEX 2 – USE CASES DESCRIPTIONS	149
UC1.POWER GRID: AI ASSISTANT SUPPORTING HUMAN OPERATORS’ DECISION-MAKING IN MANAGING POWER GRID CONGESTION	149
UC2.POWER GRID: SIM2REAL, TRANSFER AI-ASSISTANT FROM SIMULATION TO REAL-WORLD OPERATION	178
UC1.RAILWAY: AUTOMATED RE-SCHEDULING IN RAILWAY OPERATIONS	199
UC2.RAILWAY: AI-ASSISTED HUMAN RE-SCHEDULING IN RAILWAY OPERATIONS	215
UC1.ATM: AIRSPACE SECTORISATION ASSISTANT	236
UC2.ATM: FLOW & AIRSPACE MANAGEMENT ASSISTANT	250
ANNEX 3 – RELEVANT ALTAI REQUIREMENTS	266
POWER GRID	266



RAILWAY _____	270
AIR TRAFFIC MANAGEMENT _____	274
ANNEX 4 – CONTEXT, CHARACTERISTICS, IMPACT AND EVALUATION OF DECISIONS _____	278
WORD ANALYSIS _____	278
DETAILED ANSWERS _____	282
ANALYSIS OF DECISION-MAKING SCENARIO _____	327

# LIST OF FIGURES

FIGURE 1 – USE CASE TEMPLATE AND METHODOLOGY .....	19
FIGURE 2 – AI4REALNET USE CASES OVERVIEW .....	23
FIGURE 3 – ALTAI STRUCTURE.....	46
FIGURE 4 – PROCESS FOLLOWED BY AI4REALNET TO DERIVE NON-FUNCTIONAL REQUIREMENTS FROM ALTAI.....	47
FIGURE 5 – POWER GRID: RELEVANT ALTAI REQUIREMENTS.....	48
FIGURE 6 – RAILWAY: ALTAI REQUIREMENTS RELEVANT FOR POC PLANNED FOR AI4REALNET .....	50
FIGURE 7 – RAILWAY: ALTAI REQUIREMENTS RELEVANT FOR POC AND EXTENDED VERSION FOR THE REAL-LIFE APPLICATION.....	51
FIGURE 8 – ATM: RELEVANT ALTAI REQUIREMENTS.....	52
FIGURE 9 – GENERIC VIEW OF CONCEPTUAL FRAMEWORK BUILDING BLOCKS AND SECTION ORGANIZATION .....	57
FIGURE 10 – DECISIONS IN CRITICAL NETWORK INFRASTRUCTURE OPERATIONS .....	58
FIGURE 11 – DECISIONS ANALYSIS OF CRITICAL NETWORK INFRASTRUCTURES .....	59
FIGURE 12 – DETAIL OF DECISION MAKING .....	60
FIGURE 13 – LEVELS OF HUMAN-MACHINE COMPATIBILITY AND THEIR RESPECTIVE CONSTRUCTS FOUND IN COGNITIVE ENGINEERING RESEARCH ARE ORDERED BY INCREASED LEVELS OF COGNITIVE WORK; ADAPTED FROM (WESTIN ET AL., 2016) .....	77
FIGURE 14 – TRIADIC APPROACH TO HUMAN-MACHINE INTERACTION. ....	78
FIGURE 15 – MERGER OF JCF AND EID ON A FUNCTIONAL LEVEL. ....	79
FIGURE 16 – STAGES AND LEVELS OF AUTOMATION MODELLED AFTER HUMAN INFORMATION PROCESSING STEPS (PARASURAMAN ET AL., 2000).....	81
FIGURE 17 – ABSTRACT STATE-ACTION SPACE DESCRIBING A GENERIC PLANNING PROBLEM WHERE HUMANS AND AUTOMATION CAN COLLABORATE (IN SERIAL OR PARALLEL) TO BRING THE SYSTEM FROM AN INITIAL STATE TOWARD A SAFE TARGET STATE (VAN PAASSEN ET AL., 2018). ....	82
FIGURE 18 – LACC-LOA MATRIX FOR THE EXAMPLE IN FIGURE 17.....	84
FIGURE 19 – JCF SCORE FOR SCENARIO ③ .....	85
FIGURE 20 – CONCEPT OF RESILIENCE QUANTIFICATION IN TRAINING-TIME AND TEST-TIME PHASE	89
FIGURE 21 – PROTOTYPE SCHEMATIC OF A DEFERRAL MECHANISM THAT LEARNS TO DEFER DECISION-MAKING FROM THE AI MODEL TO A HUMAN.....	96

FIGURE 22 – DESCRIPTIVE SCHEMATIC OF A CO-LEARNING AI AGENT..... 97

FIGURE 23 – EXAMPLE OF MULTI-OBJECTIVE VISUALIZATION ..... 100

FIGURE 24 – FROM SUPERVISION TO HYPERVISION..... 101

FIGURE 25 – HYPERVISION IMPLEMENTATION..... 101

FIGURE 26 – EXAMPLE OF HYPERVISION INTERFACE (CAB PROJECT)..... 102

FIGURE 27 - EXAMPLE OF HYPERVISION INTERFACE (OPERATORFABRIC)..... 103

FIGURE 28 – GENERAL VIEW OF THE METHODOLOGY OF THE CONCEPTUAL FRAMEWORK ..... 104

FIGURE 29 – STAKEHOLDERS DIAGRAM..... 105

**FIGURE 30 – ENVIRONMENT DIAGRAM..... 106**

FIGURE 31 – OPERATIONAL USE CASES DIAGRAM..... 111

FIGURE 32 – ABSTRACT BASE USER STORY ..... 112

FIGURE 33 – FUNCTIONAL DECOMPOSITION..... 112

FIGURE 34 – FUNCTIONAL INTERACTION DIAGRAM..... 114

FIGURE 35 – LOGICAL ARCHITECTURE (HUMAN IN FULL CONTROL SCENARIO)..... 115

FIGURE 36 – LOGICAL ARCHITECTURE (HUMAN-AI CO-LEARNING SCENARIO) ..... 116

FIGURE 37 – LOGICAL ARCHITECTURE (AUTONOMOUS AI SCENARIO) ..... 117

FIGURE 38 – HIERARCHICAL REPRESENTATION OF THE SYSTEMS’ BUILDING BLOCKS AND CONTEXT ..... 118

FIGURE 39 – H-H VERSUS H-AI TRUST ..... 122

FIGURE 40 – AI-RELATED RISK AND ITS COMPONENTS..... 124

FIGURE 41 - AI DECISION EXPLORATION STEPS EXAMPLE ..... 327

FIGURE 42 - COMMON ANALYSIS OF DECISION-MAKING SCENARIOS ..... 328

# LIST OF TABLES

TABLE 1 – CHARACTERISTICS OF THE ENVIRONMENT ASSOCIATED WITH THE USE CASES.....	31
TABLE 2 – SUMMARY OF THE SYSTEM THREATS AND VULNERABILITIES RELATED TO THE USE CASES	33
TABLE 3 – SUMMARY OF SOCIETAL CONCERNS ABOUT THE USE CASES.....	34
TABLE 4 – LIST OF KPIS PER USE CASE .....	44
TABLE 5 – SIMILARITY SCORE OF DECISION ANALYSIS ACROSS DOMAINS .....	60
TABLE 6 – SIMILAR DECISION CHARACTERISTICS ACROSS ALL DOMAINS.....	61
TABLE 7 – TYPE OF XAI RELATED TO MACROCOGNITION AND COGNITIVE BIASES .....	65
TABLE 8 – TYPE OF XAI RELATED TO THE CRITICAL PSYCHOLOGICAL STATES AND THEIR EXPRESSION	69
TABLE 9 – TYPE OF XAI RELATED TO THE CRITICAL PSYCHOLOGICAL STATES AND THEIR EXPRESSION	73
TABLE 10 – TYPE OF XAI RELATED TO THE CRITICAL REQUIREMENTS AND THEIR EXPRESSION FOR ESTABLISHING APPROPRIATE TRUST IN AI .....	75
TABLE 11 – AN EXAMPLE OF RISK QUALITATIVE ASSESSMENT OF THE UC BASED ON THE DIMENSIONS OF ETSI GR SAI .....	87
TABLE 12 – CATEGORIES FOR THE THREE DOMAINS .....	108
TABLE 13 – HUMAN-IN-THE-LOOP AND OVERSIGHT REQUIREMENTS .....	110
TABLE 14 – SUMMARY OF THE KEY REQUIREMENTS DERIVED FROM THE ALTAI FRAMEWORK AND ADAPTED FOR AI4REALNET’S SAFETY-CRITICAL SYSTEMS .....	128

## ABBREVIATIONS AND ACRONYMS

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

# 1. INTRODUCTION

Artificial Intelligence (AI) technology has the potential to enhance the flexibility and resilience of critical network infrastructures to address global challenges like climate change, energy transition, increasing demand from mobility infrastructures, and digital transformation. However, AI faces several challenges: ensuring reliability, transparency, and ethical adherence to prevent errors and adversarial attacks; managing the complexity and uncertainty from aging assets, climate change, and rising demand in energy and mobility networks; enabling effective human-AI collaboration through reciprocal learning and integration of human knowledge; and overcoming scalability issues in AI methods like reinforcement learning (RL) when applied to large-scale infrastructures.

AI4REALNET aims to create a comprehensive multidisciplinary approach by combining emerging AI algorithms, open-source AI-friendly digital environments, and socio-technical design of AI-based decision systems with human-machine interaction. This aims to enhance the real-time and predictive operation of network infrastructures. The project focuses on three critical infrastructures — electricity network, railway, and air traffic management — vital to Europe and identified as priority sectors in national AI strategies.

AI4REALNET envisions a balanced coexistence of human control and AI-based automation, divided into three levels: a) full human control (AI-assisted), b) co-learning between AI and humans, and c) trustworthy, human-certified full AI-based control. A detailed overview and discussion of the research ideas can be found in the project's position paper (Mussi et al., 2024).

Industry-relevant and domain-specific use cases drive the project activities for applying novel AI-based methods and that i) are focused on critical challenges and tasks of network operators, considering strategic long-term goals, and ii) reproduce real operating scenarios with human operators. The use case description follows the work of ISO/IEC TR 24030, which allows a formal and structured identification of functional requirements (for the AI-based decision systems and digital environments). The analysis of non-functional requirements follows the Assessment List for Trustworthy Artificial Intelligence (ALTAI) framework. Moreover, it facilitates a comprehensive understanding necessary for conducting a thorough risk assessment following the AI Act's legal requirements.

A comprehensive multi-disciplinary framework that supports the application, development, and validation of AI-based approaches within critical network infrastructures is needed to accommodate the use cases and associated decision-making processes and for the broad integration of AI in the operation tasks of critical infrastructures. The framework is built on the top of concepts such as Joint Control Framework (Lundberg and Johansson, 2021), Trustworthiness from Confiance.AI (Braunschweig et al., 2022; Gelin, 2024), and the Humane AI Ethical Framework (Dignum, 2019). Two key components are i) the conceptualization of trustworthiness and ethical foundations and ii) the analysis of human decision-making processes in real-world situations to derive qualitative descriptions of human decision-making.

This understanding will inform the development of AI-based decision systems that are robust, ethical, and effective while being sensitive to contextual factors, as well as the evaluation of social-technical performance. Potential end-users of this framework are:

- AI developers, both from industry and academia, including human factors experts (human-in-the-loop AI-based decision systems) that need to align their development work with a “real-world” implementation perspective.
- Innovation managers from critical infrastructure operators who want to ensure that the resulting system will serve their needs as well as specific functional and non-functional requirements for the products.
- Network operation managers who want to develop a strategic and long-term vision for human-AI teaming, corresponding architectures, and requirements. The framework can be followed to build systems that satisfy/serve their needs.
- Regulatory bodies from the European Union (e.g., EU AI Office, AI Advisory Forum) and industry.
- Standardization organizations that aim to standardize the application of AI across various critical infrastructures, ensuring consistency, quality, and compatibility of AI solutions.

The remainder of this report is organized as follows: Section 2 describes the UC methodology and the six industry-driven UCs covering their goals, main functional and non-functional requirements, challenges, and key performance indicators (KPIs). Section 3 presents the conceptual framework, divided into the decision-making process through the human perspective and sociotechnical system, the AI perspective and corresponding strategies and methods, and the validation of the decision-making process through the trustworthiness and ethical assessment framework. Section 4 presents the concluding remarks.

The main body of the document is complemented by several annexes: Annex 1 with use case template; Annex 2 with UCs description; Annex 4 with a summary of the ALTAI requirements, Annex 4 with context, characteristics, impacts, and evaluation of the decisions in the critical infrastructures.

## 2. USE CASES AND METHODOLOGY

The following Sections detail the context, concepts, tools, description, and benefits of the Use Case (UC) methodology from AI4REALNET applied to critical infrastructures (see Section 2.1). Initially developed for software and systems engineering in the 1980s and 1990s, the UC methodology has since been extended to business and system process modeling. It has extensively been used within several domains, such as manufacturing, smart energy grids, and mobility, among others.

A summary description of the six industry-driven UCs from the AI4REALNET project is presented in Section 2.2, and the KPIs in Section 2.3. The project adopted the [Assessment List for Trustworthy Artificial Intelligence \(ALTAI\)](#) as a comprehensive tool for self-assessment across various dimensions, and Section 2.4 described how this methodology was applied to capture non-functional requirements related to trustworthy AI in the UC design.

### 2.1 METHODOLOGY

#### 2.1.1 HISTORY OF THE METHODOLOGY

The first UCs were written in the 1980s by Ivar Jacobson, a Swedish software engineer working at Ericsson, in order to define the architecture of one of the company's information systems. Developed as part of an Object-Oriented Software Engineering method, they were initially meant to describe situations or scenarios of usage of a given system. The UC methodology was significantly enriched and developed in the 1990s and 2000s by Alistair Cockburn, especially in his book "Writing Effective Use Cases," published in 2000, and by Kurt Bittner and Ian Spence, 2003.

Originally developed as part of the *IntelliGrid* Architecture developed by the Electrical Power Research Institute (EPRI), as a means to implement the "IntelliGrid vision" of the automated, self-healing, and efficient power system of the future, the International Electrotechnical Commission (IEC) Publicly Available Specification (PAS) 62559:2008 was issued to define a methodology for power system domain experts to determine and describe their user requirements for automation systems based on their business goals. Since its release in January 2008, the use case methodology outlined in IEC PAS 62559 has seen growing adoption within standardization efforts. This led to recognizing a need for a structured framework to ensure that IEC experts could consistently present use cases. In February 2010, the IEC Standardization Management Board SG3 recommendation 7 requested the urgent delivery of a generic use case repository for all Smart Grid applications, introduced a need to transform IEC PAS 62559 to an IEC 62559 standard to support the development of an IEC use case repository and to provide support for the use case methodology in general.

In the field of AI, the International Organization for Standardization (ISO) Technical Committee ISO/IEC JTC 1/SC 42 released a comprehensive document that compiles a wide range of AI use cases spanning different domains and sectors: "ISO/IEC TR 24030:2024. Information technology. Artificial intelligence (AI) Use cases". This document serves to aid in the establishment of AI standards, fostering collaboration, and enhancing understanding of both the potential and challenges presented by AI across industries. The technical committee used a template for collecting UC descriptions based on ISO/IEC 20547-2, IEC 62559, and the Institute of Electrical and Electronics Engineers (IEEE) P7003.



The design of social-technical systems with AI technology calls for cooperation between experts from several different domains (AI, domain-specific knowledge from areas such as power systems, railway, air traffic control (ATC), social and cognitive sciences, and human-computer interaction, among others). In the development and design of these systems, adherence to standards is crucial for achieving solutions that are interoperable, safe, secure, and cost-effective. Therefore, a common methodology for UC design is required for all involved stakeholders, and it should include terminology, quality guidelines, and workflows. This is essential not only during project development but also in the process of standardization work.

### 2.1.2 DEFINITIONS

According to IEC 62559-2, a UC describes the functions of a system under discussion (SuD) in a technology-neutral way. It identifies participating actors that can, for instance, be other systems or human actors that are playing a role within a UC. It consists of a specification of a set of actions performed by a SuD that yields an observable result that is of value for one or more actors or other stakeholders of the system. In other words, it describes, in text format, how one or several actors interact within a given system to achieve goals. UCs can be specified on different levels of granularity and are, according to their level of technological abstraction and granularity, described either as business use case (i.e., describes a general requirement, idea, or concept independently from a specific technical realization like an architectural solution) or system use case (i.e., describes in detail the functionality of a business process).

In order to clearly explain the definition, it is important to further detail the different concepts used.

- An Actor can be defined as anyone or anything with behavior. It can include:
  - Roles – the external intended behavior of a business party that cannot be shared, such as network operator, service provider, or regulator.
  - Persons – examples: human operator.
  - Information Systems – examples: Supervisory Control and Data Acquisition (SCADA), transport management system.
  - Physical components – examples: energy storage, airplane, train.
- The SuD defines the scope of a UC or a set of UCs, i.e., its boundaries. In AI4REALNET, the scope of the SuD is the AI-based decision system and the human-machine interaction to be designed, i.e., it is concerned with using AI technology to achieve a specific goal (for the organization, human operator, or citizens) by complementing and augmenting human abilities.

UCs are, above all, a textual description. Existing literature on the methodology has provided several UC templates. The AI4REALNET project adapted the IEC 62559-2 standard that defines the structure of a UC template, template lists for actors and requirements, and their relation to each other. It is a standardized template for describing UCs defined for various purposes, such as use in standardization organizations for standards development or within development projects for system development. The AI4REALNET adaptation considers the version presented in ISO/IEC TR 24030 to describe AI use cases, which is also based on ISO/IEC 20547-2, IEC 62559, and IEEE P7003.

UCs can also be depicted in diagrams using modeling languages to facilitate the presentation and the validation of Use Cases. The most commonly used standard for modeling use cases is the Unified

Modeling Language (UML), a standardized general-purpose modeling language in the field of software engineering that can also be used for business modeling. The AI4REALNET decided not to use UML descriptions for the use cases, but rather to put the main focus on the textual descriptions.

### **2.1.3 USE CASES DESIGN AND WRITING PROCESS**

#### **2.1.3.1 IDENTIFICATION**

In order to identify the UCs to be described in the present deliverable, the AI4REALNET project iteratively identified and defined the tasks at the network operators that are evolving or being created with the development of AI technologies and digitalization. This work began with the use cases defined in the Description of Action of the project and the results of consortium meetings, workshops with stakeholders, public webinars (including the possibility of receiving inputs via public consultation), and an analysis of the literature related to the AI impact of the three network infrastructures.

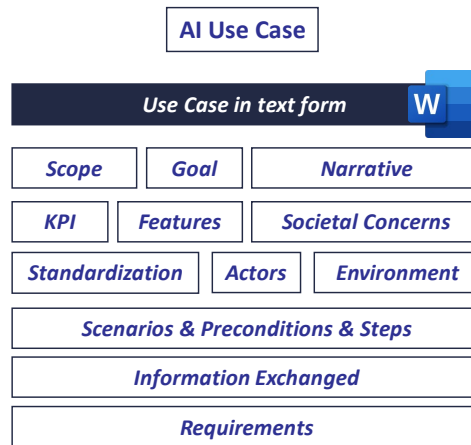
Furthermore, the network operators based this work on a thorough analysis of their roadmap, their internal organization, as well as the current regulatory framework and its evolutions in the short and medium term. This allowed them to evaluate the potential gaps to be closed to implement the identified AI-based processes and the impact on their organization. For each domain, two UCs were selected to be fully described.

#### **2.1.3.2 WRITING AND REVIEW PROCESSES**

The AI4REALNET project agreed to appoint the network operators (RTE, TenneT, DB, SBB, and NAV) as responsible for describing the selected UCs and capturing the associated functional and non-functional requirements, with the support of their domain experts and reviewed by the partner Research Institutes.

For each UC's process, the network operator responsible for describing it in UC detailed the activities of the process (or 'steps' of the Use Case) and the associated business/operational rules – including an analysis of the consistency with regional and national regulatory and legal frameworks and requirements. This work will serve as a basis for identifying the scenarios to be implemented in the digital environments of Task 1.3 and the associated functions to be developed within the project (link with WPs 1-4).

The network operators split between the identified UCs according to their respective resources allocated to the project and their domain of expertise. Each of them internally organized the requirements gathering and the UC writing processes. In particular, they focused on capturing functional and non-functional requirements and not the solutions or means required to achieve the objectives of the UC. To do so, the AI4REALNET project elaborated a Word UC template based on the template presented in ISO/IEC TR 24030 and available in Annex 1, which is summarized in Figure 1.



**FIGURE 1 – USE CASE TEMPLATE AND METHODOLOGY**

One of the most notable changes made to the ISO/IEC TR 24030 template was related to the identification of features from the real environment, the definition of operational scenarios, and the inclusion of non-functional requirements (e.g., using the ALTAI framework).

To fill the AI4REALNET UC template, each network operator started with a short version of the template (focused on the objectives, narratives, and identification of the KPIs and scenarios) and conducted interviews and workshops with the relevant experts and external stakeholders to describe the AI-based processes, their activities, and the associated information exchanges. The domain experts focused on detailing the business needs/rules and the associated functional and non-functional requirements while striving to be as generic as possible in their descriptions to avoid national or organizational specificities. To do so, the following workshops with internal (i.e., in-house experts) and external stakeholders were conducted per domain:

- Railway: 1 Feb 2024, in English with 20 participants (14 were external stakeholders), 5 Feb 2024, in German with 21 participants (10 were external stakeholders).
- Power grid: 23 Jan 2024 in English with 35 participants (29 were external stakeholders).
- ATM: 27 Mar 2024, in English with 38 participants (3 were external stakeholders).

These workshops were focused on receiving feedback for the following points:

- Relevance of the UC
- What is the role of humans in each UC?
- Risks associated
- Are relevant scenarios missing?

Finally, the research partners reviewed each of the UCs. This work allowed the person responsible for writing the UC to detail some of the identified business rules and needs further.

To validate the use cases widely and collect additional input, a public webinar was organized on 3 April 2024 (video [here](#)), and a [form](#) was created on the website for public consultation and feedback about the use cases during the project.

## 2.1.4 IDENTIFICATION OF REQUIREMENTS

UCs are designed to describe user requirements, i.e., all the functional and some of the non-functional requirements of a given system – whether it is a business process or a function.

User requirements can be defined as “the requirements of the function based on the business needs, without explicitly identifying any specific technologies or products. The same document can also cover ‘non-functional’ requirements, such as constraints, performance, security, and data interactions with other applications or systems”. In other words, they “define ‘what’ is needed without reference to any specific designs or technologies” (IEC/PAS 62559).

There are two types of requirements:

- **Functional requirements** capture the intended behavior of the system. This behavior may be expressed as services, tasks, and functions that the system must perform. Use cases are a valuable tool to capture the functional requirements of a system.
- **Non-functional requirements** capture general restrictions the system is subject to, such as pre-existing architectural constraints, architectural qualities (extensibility, flexibility, etc.), performances, reliability, and fault tolerance, among others.
  - Examples of non-functional requirements in the AI domain (Zhang et al., 2020a) include:
    - Robustness, such as fault tolerance, adaptability to data changes, and acceptable performance levels under adversarial events.
    - Efficiency, such as response times, frequency of updated results, scalability, energy consumption, and computational time.
    - Interpretability, such as the capacity to explain recommendations, adaptability to different levels of human-AI interaction, and model transparency. Note that interpretability can also be considered in the functional requirements.
    - Regulatory and legal, such as AI Act requirements, compliance with existing operational policies, and audits.
    - Security requirements include confidentiality, access restrictions, detection of failures and/or intrusions, failure management, and other safety, security, and failure issues.
    - Data management requirements include sizes, numbers of devices, amounts of data, scalability, expected growth over time, data access methods, data maintenance, and other data management considerations.
    - Interoperability issue.

As explained in Section 2.4, the **ALTAI was used to uncover additional non-functional requirements for each UC**. However, it is important to mention that the UCs do not capture all of the non-functional requirements. First of all, they do not intend to describe algorithms or aspects related to the design of a system’s user interface. Including these elements in the description only adds complexity and length to the UC, which should ideally be as simple and as concise as possible. Besides, **UCs, to be considered generic, should not be based on specific technologies, products, or solutions**.

### 2.1.5 IDENTIFICATION OF KPIS

Each use case has specific Key Performance Indicators (KPIs) that are linked to specific business/task objectives and are intended to capture the technical, economic, social, and human dimensions. These KPIs are also linked with the AI technical performance assessment functionalities embedded within existing AI-friendly digital environments (such as Grid2Op, Flatland, and BlueSky), but are not constrained by the capabilities of the digital environments and were defined to fully cover an evaluation in a real-world operational setting. Furthermore, it follows the socio-technical framework from (Weidinger et al., 2023), which, in addition to the technical components of an AI system, also considers human and systemic factors. For instance, it considers the context and interaction with the human operator.

These KPIs serve the dual purpose of measuring the effectiveness of the AI-based decision systems and identifying weaknesses and areas for enhancement. While there is no standardized list or definition, the definition of these KPIs was carefully produced by the authors of each UC using their domain knowledge, often supplemented by insights from a literature review (particularly relevant for measuring the different human-related factors).

### 2.1.6 BENEFITS OF THE METHODOLOGY FOR AI4REALNET

The UC methodology allows the representation of the characteristics of a complex system according to structuring and is, at the same time, an **iterative method**, which makes the development of UC both a science and an art (IEC 62559-2). The number of UCs initially identified and their content may vary during the writing and review process, as several UCs can be merged into one or a UC split into multiple UCs. They can also be detailed in successive steps and over a variable period of time, according to the needs and the priorities of the organization, project, or system under design. All actors can easily understand this method via a user-oriented writing style.

Secondly, the methodology is a **collective bargaining process** that is based on a pragmatic approach. It is designed to involve and actively engage different stakeholders (e.g., executives and managers, business experts and analysts, AI experts, project engineers, and policymakers) from various countries, organizations, and domains during the writing and review process. This provides an exhaustive and accurate list of requirements for the system under study and ensures that no topic or point of view has been left aside.

Thirdly, it consists of a **coherent and structured description**, which allows the analyses of key issues according to **different levels or perspectives** while ensuring global consistency:

- At a strategic level, with the identification of stances or assumptions related to the business model of a given AI-based decision system and their links with applicable regulations.
- At a business level, with the description of business processes and activities, as well as the interactions of several internal and/or external roles/systems to enable or execute them.
- At an information system level, with a detailed description of the functions supporting the business processes and the information flows they imply.

Besides, a UC can be used to analyze standards to determine whether they support the requirements described in a UC or need to be further developed to close existing gaps.

The UC methodology is particularly appropriate for describing AI-based decision systems, business processes, and functions evolving with AI technologies and digitalization, as it allows domain experts to brainstorm new requirements. On this basis, its use is relevant to identify the impact of the changes and opportunities brought by AI technologies, market development, or regulations and to answer questions such as *which existing business processes/ functions may or should evolve with AI technology?*, or *which new business processes/ functions may or should be implemented to integrate AI-based systems?*.

Finally, as discussed in (Brajovic et al., 2023), **a crucial preliminary measure for implementing a system certification (e.g., in accordance with the AI Act) involves describing a UC that provides the auditing authority with a concise overview of the current task and the AI application.** This overview facilitates the comprehensive understanding necessary to conduct a comprehensive risk assessment. Furthermore, the UC serves as a valuable tool during the development phase, ensuring adherence to compliance standards and fostering the creation of a trustworthy and resilient AI system. Notably, this practice aligns with the AI Act’s requirement for documenting Use Cases, particularly in high-risk application scenarios.

## 2.2 AI4REALNET USE CASES

### 2.2.1 OVERVIEW

Figure 2 presents an overview of the AI4REALNET UCs. The complete description of the use cases according to the project template can be found in Annex 2.





FIGURE 2 – AI4REALNET USE CASES OVERVIEW

### 2.2.1.1 POWER GRID DOMAIN

**Business problem:** Electricity networks are transforming as the ongoing decarbonization and digitalization introduce clean generation technologies, electrify demand, enable demand-side flexibility, and digitize and/or add new devices. This directly impacts supervision systems in control rooms, which have to a point where they are no longer cognitively manageable. Networks are also aging, and infrastructure developments are more limited, yet integrate more automata. AI can help to address more numerous, complex, and coordinated decisions, increasing uncertainty, overcrowded and fragmented work environments with multi-screen applications, and increasing human operator cognitive load.

**Today's operations:** Power system engineers are highly specialized, requiring thorough studies, accurate planning, and complex decision-making rather than merely following established protocols. They depend significantly on simulation tools, using both real-time and forecast data. However, they have limited access to decision-support tools like automated assistants. When faced with a problem, they manually explore solutions and verify their decisions using their simulation tools. They can adjust line connectivity on the grid to redirect power flows, modify (re-dispatch) generation levels, limit consumption by a small percentage, or use battery storage to change power flows in the electrical grid. These potential flexibilities require them to identify the most effective actions for each specific situation or context. Despite the range of options, their process relies heavily on experience and manual simulation to determine appropriate remedial measures.

**Key stakeholders:** Transmission system operators (TSOs), human operators, transmission grid users, and electricity market participants.

**UC1.Power Grid: AI assistant supporting human operators' decision-making in managing power grid congestion**

**Objectives:** The goal of a TSO, and thus human operators in the control room, is to control electricity transmission on the electrical infrastructure (transmission grid) while pursuing multiple objectives, firstly to keep the system state within acceptable limits and:

- Safely manage overloads on the electrical lines and, more specifically, remedial action recommendations;
- Make the most of the renewable energies installed by limiting the emergency redispatching call to thermal power plants emitting greenhouse gases;
- Ease the workload of the human operator needed to fulfill his/her missions;
- Integrate explainability, transparency, and trust considerations for the human operator.

**UC short description:** The AI assistant oversees the transmission grid, using SCADA data and Energy Management System tools to identify issues and categorize them for human intervention. It monitors power flows, adhering to defined operational conditions. Anticipating problems, it sends alerts to the operator with confidence levels, avoiding excessive alerts to maintain operator focus. Action recommendations include topological changes, re-dispatching, and renewable energy curtailment. The human operator selects an action or seeks more information, exploring alternatives. After the operator decides, the AI assistant provides feedback through load flow calculations and logs decisions for continuous learning and interaction improvement. This UC only addresses congestion issues, even if other types of issues can arise on the Transmission Grid and are handled by the operators (e.g., voltage values outside prescribed upper/lower limits).

**System description and role of the human operator:** This UC describes an AI assistant that provides a human operator with recommendations for actions and/or strategies, considering the abovementioned objectives. The AI assistant shall also act in a “bidirectional” manner, i.e., capitalize on the actions and the feedback from the operator with a continuous “online” learning process. Different modes of interaction between AI assistants and human operators are possible, ranging from “full human control” to “full AI control.” The selected mode depends on the industry domain and context. In this UC, an ex-ante choice is made to apply a hybrid interaction where the human operator gets the final word on AI assistant recommendations.



**Key benefits and impact of AI:** Minimize operational costs; facilitate energy transition by reducing renewable energy curtailment and improving carbon intensity of actions; reduce the workload of the human operator; increase resilience to extreme (natural and man-made) events.

**UC2.Power Grid: Sim2Real, transfer AI-assistant from simulation to real-world operation**

**Objectives:** Assess the capability of an AI assistant to be used for the operation of a “real” transmission grid, in the sense that the “real” environment does not exactly behave as the one available to the agent (that is implemented in the AI assistant) during training and simulation procedures, even if they share the same functional properties (same grid components and topology), and operational constraints. The main objectives are:

- Look at additional technical considerations to successfully deploy an AI assistant in the real world besides its sole ability to find solutions to simulated situations.
- Improving human trust when such systems are deployed in real-world environments.
- Allowing for iterative human-AI refinements with human feedback and insights.

**UC short description:** Outlines two paths for an AI assistant to manage a transmission grid. 1) In coping with real-world conditions, the AI assistant monitors grid situations, raises alerts for human intervention, and provides action recommendations, considering uncertainty from noisy and partially missing data. The human operator makes decisions based on AI suggestions, with feedback loops to continuously improve interactions and learn from realized actions. 2) When data limitations prevent full autonomy, the AI assistant alerts the human operator due to missing or poor-quality data. The operator can provide missing information to aid the AI in such cases. Enriched context, including human input and decisions, is logged for continuous learning, enhancing the AI assistant’s robustness in making recommendations for grid actions.

**System description and role of the human operator:** The AI assistant can still recommend actions to the human operator even with lower-quality data than used in training. However, this data may not enable fully autonomous recommendations, requiring the AI to seek additional feedback from the operator and raise an inaccuracy alert. When the AI cannot evaluate the need for action or a recommended action fails to produce the expected outcome, the operator can provide specific missing information to assist the AI in forecasting system states and assessing recommendations. As for the AI-assistant training, the human operator’s decision and perception will rely on “theoretical simulations” (training and simulation tools).

**Key benefits and impact of AI:** Minimize operational costs; facilitate energy transition by reducing renewable energy curtailment and improving carbon intensity of actions; reduce the workload of the human operator; increase resilience to extreme (natural and man-made) events.

**2.2.1.2 RAILWAY NETWORK DOMAIN**

**Business problem:** Growing environmental awareness and changing policies for mobility will lead to considerably more demand from railway network capacity, denser traffic, and a further need for efficiency and resilience of railway traffic management. Novel dispatching technologies or huge infrastructure investments are inevitable to maintain or improve the current quality of services. AI-based support systems can be developed to enhance dispatchers’ capabilities, aiming to automate some of today’s decision-making processes and provide support and input for human decision-making in complex operating scenarios.

**Today's operations:** In railway operations, the already densely planned schedules are disturbed by unexpected events, such as delays, infrastructure defects, or short-term maintenance. The execution of the planned timetable can only be achieved by acting on these events with frequent adaptation and re-scheduling of the planned train runs. Today, maintaining smoothly running operations requires that in operational centers, highly skilled personnel monitor the flow of traffic day and night and quickly make re-scheduling decisions. Re-scheduling measures include changing a train's speed, path, or platform. In a densely utilized railway network, local re-scheduling decisions potentially affect the entire flow of traffic, and their effect can propagate far into the future. This means that the re-scheduling task is a complex decision-making task that must integrate much context information under time constraints.

**Key stakeholders:** Railway network operators, network supervisors, railway undertaking operation managers, passengers, government, and society.

### **UC1.Railway: Automated re-scheduling in railway operations**

**Objectives:** The system's objective is to fully automate re-scheduling in railway operations to fulfill all offered services and minimize delays for the customer (passenger).

**UC short description:** Unexpected events, such as infrastructure malfunctions or delays, can occur in railway operations. In this case, the automated system must re-calculate the schedule so the requested services can be fulfilled with as little delay as possible. Adapting the schedule includes interventions, such as changing the speed curves of trains, changing the order of trains at the infrastructure element, changing the routes of trains, or changing the platform of a commercial stop at a station. An automated AI-based system is designed to manage and optimize railway schedules in real-time, ensuring efficient rail network use while minimizing passenger delays. The system is constantly monitored by a human operator who can adjust the system's configuration and identify the need for adaptation and re-training.

**System description and role of the human operator:** An AI-based re-scheduling system performs the re-scheduling task in a highly automated manner. This system observes the real-time state of all the trains and tracks in the control area of interest and automatically detects the need to intervene, decides on an intervention, and executes this intervention. Such an AI system for highly automated re-scheduling in operations is something new and unusual. The approach followed here is a first step towards introducing such a system. The highly automated AI system is treated as a new tool that is supervised and evaluated by an expert. In operations, the AI system re-schedules in a fully automated manner while the human supervisor monitors:

- The system's state in operations (e.g., number of trains, potential bottleneck in current and planned network usage)
- KPIs for the actual situations (e.g., current delay)
- Confidence/certainty of the AI system
- Intensity of intervention (how much changes to the current operational plan did the AI perform, e.g., change platform)
- The supervisor uses this information to:
  - Decide at which point it would be advisable to switch off the AI system and take over control.
  - Decide to re-configure/adjust the system in operations.

**Key benefits and impact:** Improve punctuality of trains; increase the speed in response to disruptions or changes; better use of the available capacity in the railway network.

### **UC2.Railway: AI-assisted human re-scheduling in railway operations**

**Objectives:** Aims to use AI-based methods to assist the human dispatcher in railway operations in re-scheduling train runs to fulfill all offered services and minimize delays for the customer (passenger).

**UC short description:** An AI-assistant system supports the human dispatcher. This system receives the real-time state of all the trains and tracks in the dispatcher's control area and derives possible dispatching options in case of deviations from the pre-planned schedule due to disruptions or delays. The options are presented in near real-time to the dispatcher and consist of actions the dispatcher can perform to bring the trains back or close to their pre-planned schedules. At any time during operations, the human-AI team can detect an emerging deviation of the actual state of the system from the planned state. The re-scheduling process can be initiated by various triggers such as infrastructure changes, train delays, equipment malfunctions, or potential future issues. The system is designed to detect these deviations in real-time and assess their impact on the overall schedule. The system also predicts issues that might become relevant in the future. The human learning process (e.g., to detect emerging deviations or to develop solutions) is explicitly supported by human-AI interaction.

**System description and role of the human operator:** The human provides feedback (e.g., context unknown to the system), which is used by the AI to adapt the solutions. The human agent can choose to select one of the suggestions provided by the AI systems, initiate a new solution search, or choose their own course of action. Alternatively, humans formulate a hypothesis, and the AI system provides evidence for and against these hypotheses. Moreover, a human supervisor reviews the system's performance, analyzing how effectively it responded to deviations and the impact on service delivery. Based on this review, adjustments are made to the system's parameters, such as altering the prioritization criteria, adjusting acceptable delay thresholds, or refining the algorithm for schedule recalculations.

**Key benefits and impact:** Improve punctuality of trains; increase the speed of response to disruptions or changes; better use of the available capacity in the railway network.

#### **2.2.1.3 AIR TRAFFIC MANAGEMENT DOMAIN**

**Business problem:** Air traffic density in European airspaces is steadily increasing. At the same time, pressing economic and environmental concerns force a fundamental shift towards time- and trajectory-based air traffic operations. Taken together, increased traffic loads and operational complexities may eventually drive the workload peaks of the tactical air traffic controller (ATCO) beyond acceptable thresholds, threatening the overall safety of the ATM system and hindering a smooth transition toward a sustainable future of ATM. Furthermore, for instance, in the Lisbon Flight Information Region (FIR), serviced by NAV Portugal, operational complexities arise from the activation of military areas, which can significantly restrict the usage of the upper airspace for General Air Traffic, requiring traffic to deviate horizontally, especially when in combination with unexpected events.

**Today's operations:** Today, sectorization is the sole responsibility of the ATC supervisor, who exclusively decides when and how to split and merge sectors best, warranted by situational demands and available ATCO personnel. Only scattered information is available on different platforms to aid ATC supervisors in this task. Still, there is no traffic pre-analysis tool and/or integrated decision-support

system to assist in, or even fully automate, the structuring of sectors with trajectory efficient routes (e.g., flight time and fuel burn) and sectorizations to keep the workload of the ATCO within acceptable thresholds, i.e., without exceeding sector capacity limits.

**Key stakeholders:** ATC and Flow Management Position (FMP) staff manager/supervisor, air navigation service provider (ANSP) responsible for the flight information region, tactical air traffic controller, airlines, and pilots.

#### **UC1.ATM: Airspace sectorization assistant**

**Objectives:** To partially and fully automate the sectorization 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.

**UC short description:** At ATC Centers, an operational supervisor exclusively decides when and how to split and merge sectors best, warranted by situational demands and available ATCO personnel. The degrees of freedom in sectorization involve considering horizontal (2D geometry) and/or vertical (altitude) constraints and can thus result in sectors split horizontally and/or vertically. Under nominal conditions, the supervisor typically can install several pre-fab sectorization options. However, unexpected events, such as deteriorated weather conditions, flight emergencies (e.g., aircraft equipment failure), and unscheduled ATC personnel shortages (e.g., due to sickness), may require non-standard sectorizations to be installed. An AI assistant, capable of operating under various levels of automation, will provide recommendations or even execute decisions on splitting the sector best horizontally, vertically, or both to balance the ATCO workload while ensuring safety and efficient traffic flows. It will also act bidirectionally by allowing the human operator to nudge the AI-generated recommendations in more favorable directions.

**System description and role of the human operator:** The system automatically observes the real-time data from all relevant ATM platforms, predicts how and when to sectorize, and implements prediction results either as recommendations (to the human supervisor) or automatically installs the sectorization plan. The AI system can be considered a new tool supervised and evaluated by a human expert. The AI system communicates its decisions on an auxiliary display that, for example, visualizes sector configurations on a map-like interface. At lower levels of automation, the role of the human operator (here, the ATC supervisor) is to evaluate the AI-based recommendations by requesting additional information and explanations, accepting or rejecting advisories, and nudging AI decisions in a different direction by manual interventions. All decisions and interactions will be logged, allowing the AI system to learn from human preferences continuously. At higher levels of automation, the AI recommendations are executed based on “management by consent” (= AI implements only when the human accepts) or “management by exception” (= AI implements, unless the human vetoes). At the highest level of automation, the AI system is automatically implemented, and humans can only revise the system’s decisions afterward.

**Key benefits and impact of AI:** Facilitate continuing growth of air traffic demand while maintaining high safety. Improve predictability of a certain sectorization over a certain time horizon.

#### **UC2.ATM: Flow & airspace management assistant**

**Objectives:** The system's objective is related to the flight execution phase when a military area is activated, and the ATC must issue deviations to avoid the activated area. The goal is to recommend

deviations with better sector capacity adherence and performance measured by an indicator of the environmental area – *en-route flight inefficiency of the actual trajectory*. The UC also considers the need to review the sectorization plan due to the activation of military areas and the required trajectory-efficient deviations.

**UC short description:** Some airports' activation/deactivation of military airspace can induce deviations from the flight plan routes. In this sense, to optimize the lateral deviation of the flights due to avoidance of an eventual temporary military-activated area, an AI assistant can analyze and suggest a decision in sectorization and routing of the main flows in the FIR. Human operators, more specifically the ATC and FMP supervisors, will be supported by an AI assistant in determining how to configure airspace sectors best and optimize the routes for traffic flows in the en-route sectors of the FIR. The AI assistant will also act bidirectionally by allowing the human operator to nudge the AI-generated recommendations in more favorable/acceptable directions. The airspace sectorization and flow structures, as devised by the AI and nudged by the operators in the pre-tactical phase, will be used by the tactical ATCO to manage traffic around the military-activated areas.

**System description and role of the human operator:** An AI-based system highly automates the airspace design for capacity and flow management for operational scenarios. This system automatically observes data from all relevant ATM platforms, predicts how to organize the airspace regarding routings and sectorization, and implements results as recommendations to the human operator (e.g., ATC and FMP supervisors). The AI system can be considered a new tool that is supervised and evaluated by a human expert. The AI system communicates its decisions on an auxiliary display that, for example, visualizes airspace configurations on a map-like interface. The role of the human operator (here, the ATC and FMP supervisors) is to evaluate the AI-based recommendations by requesting additional information or explanations, accepting or rejecting advisories, and nudging AI decisions in a different direction through manual interventions. All decisions and interactions will be logged, allowing the AI system to continuously learn from human preferences.

**Key benefits and impact of AI:** Facilitate continuing growth of air traffic demand while maintaining high safety. Improve a key performance environment indicator based on actual trajectory, measuring the average en-route additional distance concerning the great circle distance.

## 2.2.2 CROSS-DOMAIN ASPECTS

Table 1 presents a summary of the main characteristics of the real environment/problem associated with the UCs. This table shows common features of the environments and decision processes in the three domains, the most notable being very large observation and action spaces, mixed action types (discrete and continuous), sequential decision processes, stochastic environments with a strong dependency on weather conditions, and the AI system shall address unplanned events.

The relevant definitions and nomenclature<sup>1</sup> for this table are the following:

- *Fully observable or partially observable:* When an agent can perceive all relevant information to make decisions at any time, it is said to be fully observable. Otherwise, it is partially observable.

---

<sup>1</sup> Based on: <https://www.geeksforgeeks.org/types-of-environments-in-ai/> (accessed on June 2024)

- *Episodic or sequential:* In an episodic task environment, the agent’s actions are divided into atomic incidents or episodes. There is no dependency between current and previous incidents. In each incident, an agent receives input from the environment and performs the corresponding action. In a sequential environment, the previous decisions can affect all future decisions. The agent’s next action depends on what action it has taken previously and what action it is supposed to take in the future.
- *Deterministic or stochastic:* A deterministic system is one in which the outcomes are precisely determined through known relationships among the states and events, without any randomness. A stochastic system is one where the process's randomness and unpredictability are inherent. In such systems, outcomes are influenced by random variables and probabilities.

Feature	Power grid	Railway	ATM
<b>Observation space</b>	<ul style="list-style-type: none"> <li>▪ Partially observable</li> <li>▪ Real-time data update</li> <li>▪ Very large size, e.g., a network with around 100 nodes has more than 4,000 dimensions <i>For instance, RTE’s grid is composed of more than 25 000 nodes and 10 000 lines.</i></li> </ul>	<ul style="list-style-type: none"> <li>▪ Partially observable with limitations due to the unpredictable duration of delays and malfunctions</li> <li>▪ Real-time data update</li> <li>▪ Very large size, e.g., &gt; 10,000 trains (per day), &gt; 32,000 signals, &gt; 14,000 switches in the Swiss rail network</li> </ul>	<ul style="list-style-type: none"> <li>▪ Partially observable</li> <li>▪ Real-time data update</li> <li>▪ Very large size, e.g., &gt; 2000 flights per day, &gt; 10 observable states per flight, &gt; 8 en-route sectors, &gt; 20 coordination points per sector</li> </ul>
<b>Action Space</b>	<ul style="list-style-type: none"> <li>▪ Mixed actions: discrete &amp; continuous</li> <li>▪ Very large size: e.g., for a network with around 100 nodes, it has &gt; 65,000 different discrete actions &amp; &gt; 200 continuous actions <i>For instance, RTE’s grid is composed of more than 25 000 nodes and 10 000 lines.</i></li> <li>▪ Time horizon: intraday, meaning not more than a 24-hour forecast period</li> </ul>	<ul style="list-style-type: none"> <li>▪ Mixed actions: discrete &amp; continuous</li> <li>▪ Very large size: While the solution space grows exponentially, the action space grows linearly with the number of trains</li> <li>▪ Time horizon: typically from a few minutes to a couple of hours</li> </ul>	<ul style="list-style-type: none"> <li>▪ Mixed actions: discrete &amp; continuous</li> <li>▪ The action space of the human ATC staff manager is limited by the number of available sectors to choose from and depends on ATCO staff availability and the number of flights in the sector</li> <li>▪ Time horizon: range typically from a few minutes to a couple of hours (= pre-tactical operations)</li> </ul>
<b>Type of task</b>	Sequential	Sequential	Sequential
<b>Source of uncertainty</b>	Stochastic: weather-driven (e.g., load consumption and renewable energy generation), unplanned outages, missing or erroneous data.	Stochastic: weather-driven (e.g., the friction of wheels on rails), travel demand, disruptions (e.g., locomotives or another rolling stock issue), sensors, and communication level.	Stochastic: weather-driven, variability in traffic load, unpredicted ATCO staff shortage, variability in opening and closing military areas

Feature	Power grid	Railway	ATM
<b>Environment model availability</b>	Physical laws of the electrical grid	Although a good approximation of it can be achieved as the basic laws of physics are defined and clear, a model of the environment will be simplified in general	Aircraft performance models, International Standard Atmosphere

**TABLE 1 – CHARACTERISTICS OF THE ENVIRONMENT ASSOCIATED WITH THE USE CASES**

Regarding functional requirements in the UCs, the cross-cutting aspects described below should be emphasized.

The AI-based systems raise alerts based on their confidence level, reflecting the AI’s certainty, to ensure timely human intervention. The systems manage the alert frequency and avoid alert overload to prevent operator fatigue and maintain focus. These systems also allow for human override. For instance, in UC1.Railway, which involves automated train rescheduling, the human supervisor can decide when to switch off the AI system and take control. They can also use AI confidence levels to re-configure or adjust operational settings. Information about epistemic uncertainty can identify states worth exploring to understand the environment better or detect out-of-distribution environments (Charpentier et al., 2022).

By providing this additional layer of information, the AI helps human operators and supervisors make more informed decisions. This functional requirement aligns with the AI Act<sup>2</sup>, Article 14, “Human oversight”, in particular, “to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system”, and “to intervene in the operation of the high-risk AI system or interrupt the system through a ‘stop’ button or a similar procedure that allows the system to come to a halt in a safe state”.

The co-learning process between humans and AI enables humans to a) request additional information and explanations, accept or reject advisories, and influence AI decisions through manual interventions, and b) log all decisions and interactions, allowing the AI system to continuously learn from human preferences. In this collaborative setting, humans can also formulate hypotheses, with the AI system providing evidence for and against these hypotheses. This functional requirement is common across the three domains. It maintains human involvement by creating a feedback loop, ensuring that potentially biased outputs are addressed with appropriate mitigation measures, as Article 15 of the AI Act mentions. This approach is especially relevant in situations with incomplete information or new contexts, where the AI adapts its solutions based on human feedback or where humans provide specific missing information to help the AI forecast system states and assess action recommendations.

The AI system should be capable of providing recommendations and decisions to support the real-time operation of network infrastructures. It should integrate information and forecasted conditions to enable corrective and preventive actions at various levels of automation (Nylin et al., 2022). This allows for adjustments to the automation settings of the AI system, with the human operator's role ranging from manually implementing actions while being supported and advised by the AI system to revising

<sup>2</sup> [https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138\\_EN.pdf](https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf)

AI-implemented plans. In the power grid domain, the focus is primarily on manual actions performed by humans (that follow a common human-AI system decision process). In contrast, in the railway and ATM domains, higher levels of automation are considered for certain UCs.

Lastly, all the domains have a network structure that can provide constraints on solutions but also help inform solution strategies.

### 2.2.3 SYSTEM THREATS AND VULNERABILITIES

The three domains are susceptible to disruptions caused by unexpected events like extreme weather, technical failures, or even human resource limitations such as staffing shortages in the ATM domain. Reliable data is essential for effective decision-making, and issues like communication noise in railway systems or forecast errors in power grids can negatively impact operations.

Security threats, which include malicious actors and adversarial data attacks, and progressive deviation of environment, are a concern for both railway and power grids, where malicious actors could target the AI system to cause delays or malfunctions. To adapt to evolving environments with changing regulations, human behavior, and operational realities, all three domains require regular updates to ensure that AI-based systems remain effective.

Table 2 presents a summary of the system threats and vulnerabilities; a more detailed description can be found in Annex 2.

Feature	Power grid	Railway	ATM
<b>Trust from operators</b>	Introduce a negative cognitive bias in humans due to imperfect AI performance; accountability of decisions	Introduce a negative cognitive bias in humans due to imperfect AI performance; accountability of decisions	Introduce a negative cognitive bias in humans due to imperfect AI performance; accountability of decisions
<b>Unexpected events</b>	Weather, impact of planned maintenance, equipment failures, cyber-attacks	Weather, emergencies, staffing shortages	Weather events, flight emergencies, unscheduled ATC personnel shortages
<b>Data quality</b>	Communication/sensor noise, forecast errors	Delays; scattered information	Information scattered over various ATM systems; delayed and uncertain information
<b>Security</b>	Disruption or manipulation of the AI system, e.g., input (observation) data	Privacy and data protection; understanding failure modes, and resilience to adversarial attacks	Integrity and confidentiality of sensitive operational data; adversarial data attacks
<b>Progressive environment change</b>	System conditions evolve, but also the operational rules, the human operators' behavior, or regulations	Shift in skills for human operators; system conditions evolve, and operational rules	AI needs regular updates to adapt to changing conditions



Feature	Power grid	Railway	ATM
<b>Mismatch between training &amp; deployment</b>	Ineffective control actions to solve congestion problems; expensive control actions and excessive curtailment of renewables	Decrease in the trustworthiness of the railway operator; introduce inequality in service quality for different geographic regions	Inaccurate assumptions about real-world conditions; updated information deviates from the information/data used for the implemented sector plan

TABLE 2 – SUMMARY OF THE SYSTEM THREATS AND VULNERABILITIES RELATED TO THE USE CASES

## 2.2.4 SOCIETAL CONCERNS

Safety is a major focus for all three sectors. Power grids have additional concerns around integrating renewable energy sources and maintaining resilience against extreme events and cyberattacks. Public trust, data privacy, and clear accountability are key concerns for AI for the three sectors, and all face potential job displacement anxieties due to automation. Table 3 presents a summary of the societal concerns; a more detailed description can be found in Annex 2.

Feature	Power grid	Railway	ATM
<b>Main driver(s)</b>	Enable higher integration levels of renewable energy and decarbonization of the economy while maintaining (or improving) the reliability and resilience	Traffic density on the European rail networks is constantly increasing; densely planned schedules are disturbed by unexpected events (e.g., infrastructure defects, delays).	Maintaining safe and efficient ATM under increased traffic loads while adhering to the workload capacity limits of tactical ATCOs
<b>Privacy &amp; data protection</b>	Data storage, processing, security	Data storage, processing, security (GDPR compliance)	Secure handling, storage, and processing of sensitive information
<b>Transparency &amp; accountability</b>	Human operators shall be able to understand the ground basis of AI action recommendations	Concerns about AI decision-making and accountability for failures	Explainability of AI recommendations, operator oversight
<b>Employment &amp; skill shift</b>	Human operator’s sole ability to operate the grid and associated knowledge shall not be hampered by the AI system	Potential job displacement and the need for staff reskilling	Job displacement and the need for reskilling of ATC staff
<b>Public trust &amp; acceptance</b>	External supervision and regulator conformity assessment are present	Risk of severe traffic congestion with significant economic effects on the network in case of a malfunctioning AI	Apprehensions and resistance from the public regarding the shift to AI-driven systems

Feature	Power grid	Railway	ATM
<b>Safety &amp; security</b>	Failure modes, model robustness, preventing adversarial attacks; avoiding propagation to other critical infrastructures	Maintain robust data protection and cybersecurity measures	System performance under extreme events, cybersecurity concerns
<b>Inequality</b>	Risk of unequal service quality due to AI bias	Inequality in service quality for different geographic regions or categories of passengers	Disparities in service quality (potentially favoring certain airspaces, airlines, or regions over others)

TABLE 3 – SUMMARY OF SOCIETAL CONCERNS ABOUT THE USE CASES

## 2.2.5 STANDARDIZATION OPPORTUNITIES

Standardization in AI for critical infrastructures is fundamental to ensure reliable and secure implementation of AI-based decision systems, enhancing interoperability while mitigating risks and safeguarding essential services and legacy systems. While contributions to standards are beyond the scope of the AI4REALNET project, the use case descriptions also serve as a tool to standardize processes and identify potential standardization opportunities. Therefore, this subsection summarizes relevant existing standards and opportunities identified in the use case descriptions of Annex 2.

The following existing standards were considered relevant for the use cases across all three domains:

- *ISO/IEC 23894:2023, Information technology — Artificial intelligence — Guidance on risk management.* All UCs involve high-stakes tasks, and thus, risk management specifically related to AI is fundamental. This standard describes the principles applied to AI, risk management framework, and processes.
- *ISO/IEC 38507:2022, Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations.* AI assistants, co-learning systems, and fully autonomous AI require an analysis of the governance implications associated with their use. This includes studying data-driven problem-solving and adaptive AI systems, such as retraining during the operational phase, to adapt to new operating conditions and/or human feedback, culture, and values about stakeholders, markets, and regulation.
- *ISO/IEC 42001:2023, Information technology – Artificial intelligence – Management system.* For organizations, it sets out a structured way to manage risks and opportunities associated with AI, balancing innovation with governance.
- *IEEE 7000-2021, IEEE Standard Model Process for Addressing Ethical Concerns during System Design.* Defines a framework for organizations to embed ethical considerations in concept exploration and development. It promotes collaboration between key stakeholders and ensures ethical values are traceable throughout the design process, impacting the operational concept, value propositions, and risk management.

### Relevant standards and standardization requirements in the power grid domain

As highlighted in an ENTSO-E report (ENTSO-E, 2019), additional interoperability and standards are crucial for the cyber-physical system supporting the energy transition. This need has become even

more urgent due to the AI Act requirements for the energy sector. This will further require intensifying the standardization activities in AI safety and liability toward a standard definition of AI compliance requirements, test protocols, and accountability (Heymann et al., 2023).

One requirement identified in the power grid UCs is the application of an ontology that leverages agent-oriented AI recommendations to aid power grid operators in solving future problems based on past observations stored in a knowledge database. The French project Cockpit and Bidirectional Assistant (CAB) initiated the first work in this direction (Amdouni et al., 2023). Note that in other domains of the energy sector, a good example of the use of ontologies is the Smart Applications REference (SAREF) ontology, a family of standards that enables interoperability between solutions from different providers and among various activity sectors on the Internet of Things and therefore contributes to the development of the global digital market. A similar initiative should be promoted for AI assistants.

Another emerging trend, already aligned with standardization initiatives like ISO/IEC 24029-2:2023, is formal verification methods for artificial neural networks (Venzke and Chatzivasileiadis, 2021). These methods can help estimate the operating boundaries of AI systems and provide mathematical guarantees, which are crucial for their deployment in critical applications. However, these standards should go beyond artificial neural networks and consider other AI models, as well as the communication of this information to the end-user/decision-maker and the interaction between AI and the environment.

The Testing and Experimentation Facilities from Horizon Europe for the energy domain will play an important role in the standardization of conformal verification methods and in AI testing across the technology readiness level chain (Cremer et al., 2024).

### **Relevant standards and standardization requirements in the railway network domain**

In railways, there are different levels of automation (Grade of Automation, GoA) defined in the IEC 62267 Standard (*Railway applications - Automated urban guided transport (AUGT) - Safety requirements*). This standard covers high-level safety requirements applicable to automated urban guided transport systems, with driverless or unattended self-propelled trains operating on an exclusive guideway. Furthermore, standard DIN EN 50126, *Railway Applications – The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS)*, considers the generic aspects of the RAMS life cycle and describes a safety management process. It provides guidelines for defining requirements, conducting analyses, and demonstrating the reliability, availability, maintainability, and safety aspects throughout the lifecycle of railway applications. Another standard is DIN EN 50128, *Railway applications – Communication, signaling and processing systems*, which outlines the procedural and technical criteria for crafting software intended for programmable electronic systems in railway control and protection applications. A detailed review of standards and AI for railway operations can be found in (Gesmann-Nuissl & Kunitz, 2022). This establishes that an important standardization requirement is related to AI safety requirements.

Moreover, in the AI4REALNET use cases, there are opportunities for co-decision-making and human-computer interaction. These include standardizing bidirectional communication in the decision-making process, allowing humans to use the system as a decision-support tool, and providing additional context and feedback to the AI to enhance decision-making.

### Relevant standards and standardization requirements in the ATM domain

In the ATM domain, the ICAO DOC 4444 (*Procedures for Air Navigation Services – Air Traffic Management*) is an essential document published by the International Civil Aviation Organization (ICAO) since it details the standardized procedures necessary to ensure safe, efficient, and orderly air traffic operations. Given the dynamic nature of the aviation industry, it undergoes regular updates and revisions to incorporate technological advancements, operational experiences, and emerging best practices. ICAO DOC 4444 is evolving to accommodate AI-based systems<sup>3</sup>.

In the AI4REALNET UCs, a key standardization objective is to define a uniform set of KPIs to assess the effectiveness of AI-driven sectorization systems, comparing their performance (e.g., robustness, human-acceptance) with heuristic methods in prediction and planning systems. This requires implementing standardized test procedures for evaluating AI performance, with existing procedures serving as foundational benchmarks.

## 2.3 KEY PERFORMANCE INDICATORS

Using the methodology described in 2.1.6Section 2.1.6, a set of potential KPIs were identified for each use case and are listed in Table 4 in terms of definition and calculation methodology. The list of KPIs and calculation methodology will be refined in deliverable D4.1 (WP4) considering what can be computed with the project’s digital environments.

Use Cases ID	KPI name	Definition	Calculation methodology
<b>UC1.Power Grid</b> <b>UC2.Power Grid</b>	Operation score	The operation score for operating a power grid includes the cost of a blackout <sup>4</sup> , the cost of energy losses on the grid <sup>5</sup> , and the cost of remedial actions <sup>6</sup> .	<p>To simplify the computation and without hindering future improvements, it is proposed to define it as a vector with dimensions representing different units, at least:</p> <ul style="list-style-type: none"> <li>• Number of real-time topological actions (e.g., switching actions). Only unitary actions at each timestep are considered, meaning a tuple action would be counted as two separate actions.</li> <li>• Number of redispatching actions (including but not limited to storage)</li> <li>• Sum of redispatched energy volumes</li> <li>• Number of curtailment actions</li> <li>• Sum of curtailed energy volumes</li> <li>• Electricity losses</li> </ul> <p>Further details about operation score calculation will be defined in deliverable D4.1. This score could, for example, be</p>

<sup>3</sup> European Plan for Aviation Safety 2022 – 2026: <https://www.easa.europa.eu/en/document-library/general-publications/european-plan-aviation-safety-2022-2026>

<sup>4</sup> Calculated by multiplying the remaining electricity to be supplied by the market price of electricity.

<sup>5</sup> determined by multiplying the energy volume lost due to the Joule effect by the market price of electricity.

<sup>6</sup> the sum of expenses incurred by the actions using flexibilities (e.g. balancing products, curtailment or redispatching), based on the energy volume and underlying flexibility cost.

Use Cases ID	KPI name	Definition	Calculation methodology
			completed with more financial aspects, such as immediate or long-term costs (e.g., indirect costs due to the lifetime decay of circuit breakers).
<p><b>UC1.Power Grid</b> <b>UC2.Power Grid</b></p>	Network utilization	It is based on the relative line loads of the network, indicating to what extent the network and its components are utilized	<p>This can be quantified by:</p> <ul style="list-style-type: none"> <li>For each timestamp, the highest encountered N-1 line's load and N line's load</li> <li>The average of the maximum N-1 line's load and N line's load</li> <li>For each timestamp, the number of lines where the N-1 line's load is greater than a given threshold (e.g., 1.0)</li> <li>For each timestamp, the number of lines where the N line's load is greater than a given threshold (e.g., 0.9)</li> <li>For all timestamps, the energy of overloads, calculated as the power exceeding the line capacity, integrated over the concerned timestamps (in N and N-1 state)</li> </ul>
<p><b>UC1.Power Grid</b> <b>UC2.Power Grid</b></p>	Topological action complexity	It 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.	<p>Metrics for quantifying the topological utilization of the grid:</p> <ul style="list-style-type: none"> <li>The average number of split substations (gives an indication of the distance to the reference topology)</li> <li>The average number of substations modified in one timestamp (gives an indication of the complexity of the topological actions)</li> <li>Number of unique split substations</li> </ul>
<b>UC1.Power Grid</b>	Assistant alert accuracy	It is based on the number of times the AI assistant agent is right about forecasted issues (e.g., overloads) ahead of time.	<p>Confusion matrix calculated to show:</p> <ul style="list-style-type: none"> <li>True positive cases: forecast alerts were raised by the AI assistant, and the problem did occur on the transmission grid</li> <li>False positive cases: forecast alerts were raised by the AI assistant, but no problem occurred on the transmission grid</li> <li>False negative cases: The AI assistant raised no forecast alert, but problems occurred on the transmission grid</li> </ul>
<p><b>UC1.Power Grid</b> <b>UC2.Power Grid</b> <b>UC2.Railway</b></p>	Assistant relevance	<b>Power Grid:</b> It is based on an evaluation by the human operator of the relevance of action recommendations provided by the AI assistant.	Measured by the number of recommendations from the AI assistant effectively used by the human operator. It has a range of [0, 100] where:

Use Cases ID	KPI name	Definition	Calculation methodology
		<p><b>Railway:</b> Situation awareness of the human operator using the system.</p>	<ul style="list-style-type: none"> <li>0 means that no action recommendation from the AI assistant was considered useful by the human operator</li> <li>100 means that all action recommendations from the AI assistant were considered useful by the human operator</li> </ul> <p>The KPI can have values between 0 and 100 if only a part of the action recommendations from the AI assistant were used by the human operator.</p> <p>The KPI shall distinguish between the “best decision given the information available at the time” and the “best decision in hindsight.” The evaluation shall focus on the first case, i.e., it shall not be done after the facts with full knowledge of the human operator, which was unavailable at the time.</p>
<p><b>UC1.Power Grid</b> <b>UC2.Power Grid</b></p>	Action recommendation selectivity	This KPI measures how recommended actions from AI assistants contrast among KPIs used for human decisions: this allows us to put recommended actions in perspective with trade-offs used in human decisions.	<p>For each recommended action from the AI assistant, this KPIs consists of calculating the increase of each of the following KPIs (see above) due to action implementation:</p> <ul style="list-style-type: none"> <li>Network utilization</li> <li>Topological action complexity</li> <li>Operation score</li> </ul>
<p><b>UC1.Power Grid</b> <b>UC2.Power Grid</b></p>	Assistant disturbance	It aims to measure if the AI assistant's notifications are disturbing the human operator's activity.	<p>For each notification, the score has a range of [0, 5] where:</p> <ul style="list-style-type: none"> <li>0 means that the notification was not considered disturbing at all by the human operator</li> <li>5 means that the human operator considered the notification as fully disturbing</li> </ul>
<p><b>UC1.Power Grid</b> <b>UC2.Power Grid</b></p>	Workload	It is based on the workload assessment of human operators of the AI assistant.	It shall be determined according to the NASA-TLX <sup>7</sup> methodology or similar <sup>8</sup> .
<p><b>UC1.Power Grid</b> <b>UC2.Power Grid</b></p>	Total decision time	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.

<sup>7</sup> <https://humansystems.arc.nasa.gov/groups/tlx/index.php>

<sup>8</sup> See more recent works about design recommendations to create algorithms with a positive human-agent interaction and foster a pleasant user-experience: <http://hdl.handle.net/1853/61232>

Use Cases ID	KPI name	Definition	Calculation methodology
UC1.Power Grid UC2.Power Grid	Carbon intensity	It is based on the overall carbon intensity of the action recommendation	Calculated as follows: <ul style="list-style-type: none"> <li>The amount of energy curtailed (or decreased following redispatching action) is split according to generation type with a negative sign</li> <li>The amount of additional energy yielded by redispatching action is split according to generation type with a positive sign</li> <li>The netted amount of energy <math>E_i</math> (MWh) is calculated per generation type <math>i</math></li> <li>Each amount <math>E_i</math> is multiplied by the corresponding emission factor (kgCO<sub>2</sub>/MWh) <math>F_i</math></li> <li>The score is then calculated as:                             <math display="block">\frac{\sum_i E_i \times F_i}{\sum_i E_i}</math> </li> </ul>
UC1.Power Grid UC2.Power Grid UC1.Railway UC2.Railway	Trust towards the AI tool	“(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 & Forzy, 2009, p. 1261).	The human operators' trust towards the AI tool can be measured using the Scale for XAI (Hoffman et al., 2018) or similar.
UC1.Power Grid UC2.Power Grid UC1.Railway UC2.Railway	Human motivation	“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. 54).	The human operators' perceived internal work motivation can be measured by using the Job Diagnostic Survey (Hackman & Oldham, 1974) or a similar method. The questionnaire must be adapted to the AI context (e.g., problem detection with AI assistance).
UC1.Power Grid UC2.Power Grid UC1.Railway UC2.Railway	Human control/autonomy over the process	“Autonomy is the degree to which the job provides substantial freedom, independence, and discretion to the employee in scheduling the work and in determining	The control/autonomy of the human operator over the process must actually be given. This can be measured indirectly using the Work Design Questionnaire (Morgeson & Humphrey, 2006) or similar. The questionnaire must be adapted to the AI

Use Cases ID	KPI name	Definition	Calculation methodology
		the procedures to be used in carrying it out” (Hackman & Oldham, 1975, p. 162). It consists of three interrelated aspects centered on freedom in decision-making, work methods, and work scheduling (Morgeson & Humphrey, 2006). Parker and Grote (2022) view job autonomy interchangeably with job control.	context (e.g., problem detection with AI assistance).
<b>UC1.Power Grid</b> <b>UC2.Power Grid</b> <b>UC1.Railway</b> <b>UC2.Railway</b>	Human learning	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.	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 a similar method. The questionnaire needs to be adapted to the AI context.
<b>UC1.Power Grid</b> <b>UC2.Power Grid</b> <b>UC1.Railway</b> <b>UC2.Railway</b>	Decision support for the human operator	Decision support tools should be aligned with the cognitive decision-making process that people use when making judgments and decisions in the real world and ensure that the human operator retains agency (Miller, 2023). Therefore, AI decision support tools should help people remain actively involved in the decision-making process (e.g., by helping them critique their own ideas) (Miller, 2023).	The decision support for the human operator can be measured based on the criteria for good decision support (Miller, 2023) or similar. The instrument must be further developed.
<b>UC1.Power Grid</b> <b>UC2.Power Grid</b> <b>UC1.Railway</b> <b>UC2.Railway</b>	Ability to anticipate	The ability to anticipate. Knowing what to expect, or being able to anticipate developments further into the future, such as potential	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



Use Cases ID	KPI name	Definition	Calculation methodology
		disruptions, novel demands or constraints, new opportunities, or changing operating conditions (Hollnagel, 2015, p. 4).	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., 2010) or similar. The instrument needs to be further developed and adapted to the AI context.
<b>UC1.Power Grid</b> <b>UC2.Power Grid</b> <b>UC1.Railway</b> <b>UC2.Railway</b>	Situation awareness	“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, p. 12).	The human operator’s situation awareness can be measured using the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1988) or similar.
<b>UC2.Power Grid</b>	Technical robustness to real-world imperfections	Describes the ability of the AI system to maintain its performance level under natural or adversarial perturbations, namely bad or low-quality data, or when recommended action does not have the expected impact on the transmission grid’s state	This KPI can be quantified by comparing the technical performance of the AI assistant without and with the perturbations, using KPIs from UC1.Power Grid. From those KPIs, the following metrics (or properties) can be computed: 1) The extent to which the output of the AI system or a specific KPI (e.g., operation score) varies with the perturbations, e.g., measured with the output/KPI variance and/or average difference. 2) Assess whether a particular decision holds for input variation (data quality issue) in the same context. During the training time of the AI assistant, the slope of the reward/loss function deterioration can also be used to measure technical robustness.
<b>UC2.Power Grid</b>	Resilience to real-world imperfections	Ability to prepare for and adapt to changing conditions and withstand and recover (to a “normal” state) rapidly from natural or adversarial perturbations or unexpected changes.	This KPI can be quantified with the magnitude and/or duration of reward/loss function performance degradation compared to an unperturbed system for the same context. It can, for instance, be measured by the area between the reward curves of the unperturbed and perturbed AI system. This can be computed during training or operational testing time.
<b>UC2.Power Grid</b>	Transferability across fidelity levels	Measures how effectively a policy or model trained in one environment (low-fidelity simulation) performs when applied to different environments (e.g., high-fidelity simulation or real-world operation).	Evaluated by directly applying the policy trained in a low-fidelity simulation to a high-fidelity simulation and measuring its effectiveness by computing the KPIs from UC1.Power Grid.

Use Cases ID	KPI name	Definition	Calculation methodology
UC2.Power Grid	Generalization to different grid operating conditions	The ability of a policy to perform well in an unseen grid operating condition that was not part of the training experience.	Tested by exposing the previously trained AI system to different environments with changed grid elements and observing how well it adapts and performs by determining the KPIs from UC1.Power Grid.
UC2.Power Grid	Assistant self-awareness	It is based on the number of times the AI assistant agent is right about its ability to perform action recommendations ahead of time.	Confusion matrix calculated to show: <ul style="list-style-type: none"> <li>• True positive cases: AI assistant raises inaccuracy alert indicating it has insufficient data to estimate the state of the grid and it actually does not have the required data</li> <li>• False positive cases: AI assistant raises inaccuracy alert indicating it has insufficient data to estimate the state of the grid, but it actually does have the required data (i.e., it should be confident, but it is not)</li> <li>• False negative cases: AI assistant does not raise inaccuracy alert, but in reality, it cannot properly assess the situation (i.e., is falsely confident)</li> </ul> Note: This KPI is the adaptation of the "Assistant alert accuracy" KPI of UC1 "Power Grid Assistant"
UC1.Railway UC1.ATM UC2.ATM	Acceptance score	<b>Railway:</b> Tracks the frequency of human operator interventions in AI decisions. Target: Reduce to less than x% of cases.  <b>ATM:</b> Measure of acceptance degree of the generated AI solution for human operators	<b>Railway:</b> (Number of human interventions / Total AI decision instances) x 100.  <b>ATM:</b> Reflects the acceptance choice in the AI's system decision. (0% - 100%). Measured directly from yes/no/revision input, translated into % across the operator's multiple interactions with AI-generated solutions.
UC2.Railway	Acceptance	Acceptance of the system by a human user.	Using the TAM model (technology acceptance model).
UC1.Railway UC2.Railway	Punctuality	<b>UC1:</b> Measures the percentage of trains arriving at their destinations on time. Target: Achieve a punctuality rate of x% or higher. <b>UC2:</b> An aggregated measure of the delay in a scenario (defaults to be defined).	<b>UC1:</b> (Number of on-time arrivals / Total number of arrivals) x 100.  <b>UC2:</b> Sum of individual train delays divided by number of trains.
UC1.Railway UC2.Railway	Response time	<b>UC1:</b> Assesses the speed at which the AI system responds to disruptions or changes. Target: Response within x minutes of disruption detection. <b>UC2:</b> Assesses the speed at which the AI system responds to disruptions or changes. Target: Response within x minutes of disruption detection.	<b>UC1:</b> Average time taken from disruption detection to system response.  <b>UC2:</b> Average time taken from disruption detection/prediction to suggestion of adjusted schedule(s).

Use Cases ID	KPI name	Definition	Calculation methodology
		<b>UC2:</b> The time needed to produce a new schedule in case of a disturbance event.	
<b>UC1.Railway</b>	Delay reduction efficiency	Quantifies the effectiveness of the system in reducing delays. Target: Reduce overall delays by 30%.	(Total delay duration before AI implementation - Total delay duration after AI implementation) / Total delay duration before AI implementation.
<b>UC2.Railway</b>	Human information processing	The volume of information that humans consider when making decisions with AI support (compared to making decisions with no AI support).	It is measured indirectly from user interaction with the AI system through the user interface (e.g., eye-tracking, clicks, requests for information, ...) and via questionnaires answered by the human operator after use.
<b>UC2.Railway</b>	Comprehensibility	It is defined as the ability to understand a decision logic within a model and, therefore, the ability to use this knowledge in practice (Futia and Vetrò, 2020).	Comprehensibility is derived from questionnaires answered by the human operator after use.
<b>UC1.ATM</b> <b>UC2.ATM</b>	Agreement score	Measures how much the supervisor agrees with AI-generated sectorization. <i>Note: agreement and acceptance are not the same. One can accept a solution but not necessarily agree with it. A good system fosters a high-level agreement</i>	It is measured directly from user input using an agreement rating scale of 0 – 100%.
<b>UC1.ATM</b> <b>UC2.ATM</b>	Trust in AI solutions score	How much of the operator's confidence in the AI-generated solution, with and without the need for additional explanations.	It is measured directly from user input using Likert scales.
<b>UC1.ATM</b> <b>UC2.ATM</b>	Decision support satisfaction	System effectiveness in supporting the efficient decision-making by airspace managers	It is measured directly from user input using Likert scales.
<b>UC1.ATM</b> <b>UC2.ATM</b>	Efficiency score	How many times was an AI-generated solution revised? A good system would minimize the number of human interventions.	Reflects the efficiency of the combined human-AI team performance. (0% - 100%). Measured directly from user input (was the solution modified? Yes/No), translated into % across the operator's multiple interactions with AI-generated solutions
<b>UC1.ATM</b> <b>UC2.ATM</b>	Significance of human revisions	The extent of human revisions compared to the AI decision. Here, small, localized revisions (e.g., merging two small adjacent sectors in the northeast corner of the FIR) would be rated differently from larger or multiple	Reflects the AI system performance. (LOW, MED, HIGH interaction %). Measured directly from user input (of the modified solutions, how much interaction was measured? LOW number and extent of changes, MEDIUM number, and extent of changes HIGH number and extent of changes), translated into % across the

Use Cases ID	KPI name	Definition	Calculation methodology
		revisions across various areas in the FIR.	operator's multiple interactions with AI-generated solutions
UC1.ATM UC2.ATM	System reliability	System trustworthiness - operation as expected under several conditions without major failures.	Reflects the efficiency of the combined human-AI team performance. (0%-100%). Measured directly from how many times the AI-generated solutions are sound or lead to failures.
UC1.ATM UC2.ATM	AI prediction robustness	Measure the robustness of the predicted sectorization considering small variations in factors such as time horizon or capacity.	Reflects the efficiency of the combined human-AI team performance.  Measured directly from the AI generated solutions, as the average of how big a variation in capacity has to be to cause the AI to revise its previous solutions.
UC1.ATM UC2.ATM	Prompt demand rate	Assess how many times the ATCO prompts additional explanations from the AI-generated solutions.	Reflects the AI system performance. (LOW, MED, HIGH interaction %)  Measured directly from user input (how much interaction with explanations occurred and how the generated scenario is rated using the 'dynamic density index', measuring complexity), translated into % across the operator's multiple interactions with AI-generated solutions
UC1.ATM UC2.ATM	AI co-learning capability	The capability of the AI system to adapt to human preferences as perceived by ATCOs.	Measured directly from user input using Likert scales.
UC1.ATM UC2.ATM	Human response time	Time needed to react to AI advisory/information	(LOW, MED, HIGH response time %). Measured directly from user input (dismiss a window when they feel satisfied after evaluating a scenario, LOW less than 5 min, MEDIUM 5-10 min, HIGH more than 10 minutes), translated into % across the operator's multiple interactions with AI-generated solutions.
UC2.ATM	Reduction in delay	Percentual reduction of flight delays due to AI implementation in airspace and air traffic management	0% - 100%, calculated by the additional flown track miles (in combination with flown speed and altitude profiles) relative to the shortest great circle distance (and preferred speed and altitude profiles), resulting in a percentual flight time deviation.
UC2.ATM	Workload perception	Assess ATCOs perception of the system's impact on their workload (either positive or negative)	It is measured directly from user input using a 7-point Likert scale, where 1 is a huge increase in workload, and 7 is a huge decrease in workload.

TABLE 4 – LIST OF KPIS PER USE CASE

## 2.4 ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE (ALTAI)

### 2.4.1 METHODOLOGY

At this stage, AI4REALNET uses the ALTAI assessment tool to identify the relevant risks and ethical concerns and translate them to non-functional (and functional) requirements in the UCs, in alignment with the framework for trustworthy AI (TAI) established by the high-level expert group on artificial intelligence (AI HELG) appointed by the European Commission<sup>9</sup>. This framework is also the basis for the AI Act (Fedele et al., 2024). This process also allows us to evaluate the suitability of applying ALTAI at the early stages of development, identify limitations, and provide recommendations for its improvement. It also serves as a basis for establishing improved mechanisms for continuing the trustworthiness assessment during the rest of the project.

Noteworthy, the ALTAI has been conceived as an assessment instrument for *ex-post* self-assessment of AI systems. Despite this fact, we proactively used its structure to perform an *ex-ante* assessment of the UC definition in accordance with the framework for TAI from the European Commission. This allows the consortium to

- Identify risks and ethical issues particularly relevant to the considered UCs
- Define UC requirements to be fulfilled by the solutions developed in the project
- Develop suitable metrics to validate that these requirements are appropriate and sufficient to mitigate the identified risks and ethical concerns.

It is important to note that, in complement to the analysis presented in this section, Section 3.3 of the conceptual framework establishes the foundation, from both epistemological and philosophical perspectives, for a non-calculative approach to AI risk assessment and suggests modifications to the application of ALTAI in safety-critical systems.

#### 2.4.1.1 BACKGROUND

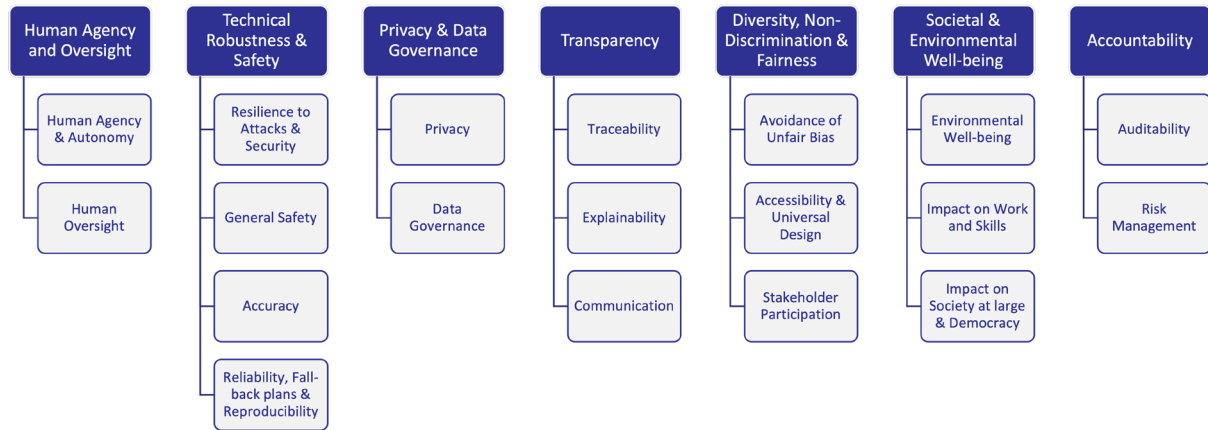
The ALTAI serves as a self-assessment checklist that aids developers in implementing key requirements according to the ethical dimensions raised therein. The ALTAI is structured in seven key ethical requirements:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination, and fairness
- Societal and environmental well-being
- Accountability

---

<sup>9</sup> <https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai>

Hereafter, we will refer to these requirements as *ethical dimensions* to avoid confusion with the UC’s requirements. The ALTAI provides a set of yes/no questions on each of these dimensions and their sub-sections to guide the self-assessment. This way, it serves as a checklist to identify issues that have not been addressed and suggests their consideration. The overall ALTAI structure is depicted in Figure 3.



**FIGURE 3 – ALTAI STRUCTURE**

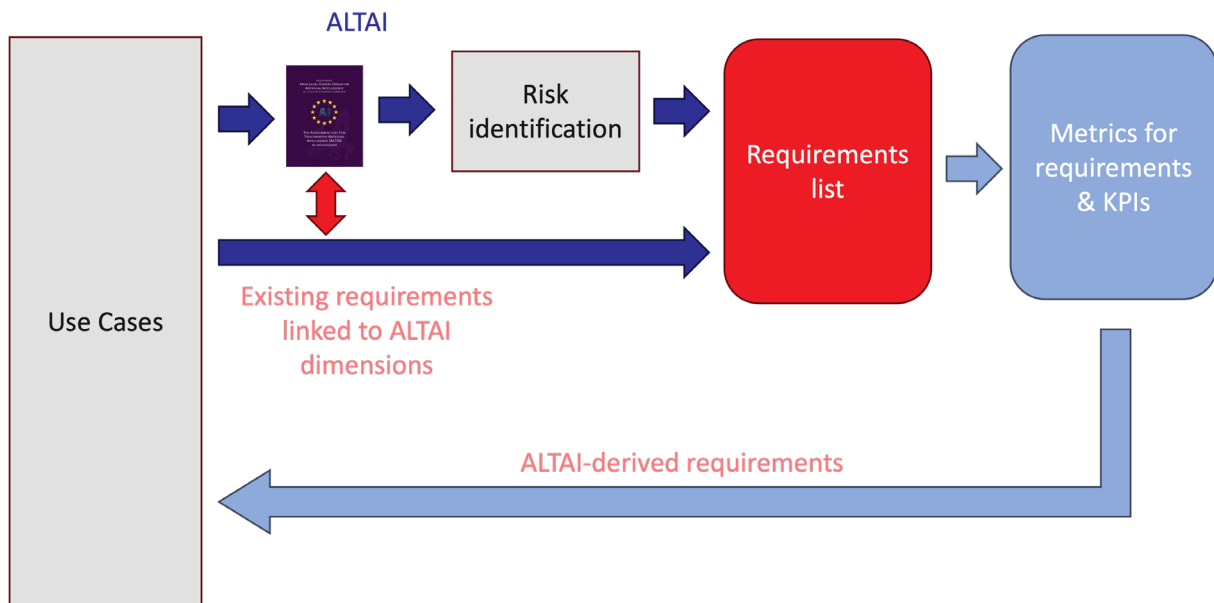
Hence, the ALTAI poses a post hoc assessment tool that identifies risks and ethical concerns. However, ALTAI does not provide:

- Means to assess if the risks have been properly identified
- Comprehensive guidance on how to address the identified risks
- Validation of whether the measures taken are appropriate and sufficient to mitigate identified risks.

There are several precedents of using ALTAI assessment for critical applications such as Advanced Driver-Assistance Systems (Borg et al., 2021), Air Traffic Controller Operations (Stefani et al., 2023) and considerations for safe autonomy of smart railways (De Donato et al., 2022). Further consideration of ALTAI is provided in (Radclyffe et al., 2023).

#### 2.4.1.2 EXTRACTING NON-FUNCTIONAL REQUIREMENTS

Figure 4 illustrates the process followed to derive requirements from the ethical assessment for the AI4REALNET UCs.



**FIGURE 4 – PROCESS FOLLOWED BY AI4REALNET TO DERIVE NON-FUNCTIONAL REQUIREMENTS FROM ALTAI**

Since each application domain has its own specific characteristics, individual assessments were pursued for the TSO, Rail, and ATM domains. The identification of requirements based on the ALTAI questionnaire followed the iterative development of the UCs (c.f. Section 2.1.3.1),

The knowledge of the domain experts is key for identifying ethical concerns. Internal workshops were held with the consortium partners involved in each application domain to introduce the ALTAI structure and the methodology. In this workshop, a first analysis of the initial versions of the use case was performed to identify aspects relating to ALTAI dimensions. These originated from the stakeholders’ interests, which are bound to reflect some individuals’ interests (and thus relate to ethical concerns) or societal concerns. This step explicitly links requirements already identified in the UC to the ALTAI structure.

After these workshops, the ALTAI questionnaire was provided as a shared document in which the workshop participants, stakeholders, developers, and all other parties involved in each UC provided answers to the individual questions. They add their insights, comments, and perspectives. One individual from the domain experts is designated to organize and manage this process. Based on these, for each question, a conscientious decision is made, which comprises:

1. A decision: Is the issue raised by the ALTA question relevant, and must it be addressed in the UC? This decision and its supporting arguments must be recorded in the ALTAI document for the UC.
2. If the issue is deemed relevant, the respective UC requirement(s) that address the issue are recorded and included in the Requirements section of the UC template – See Annex 1.

The argumentation regarding the relevance of the ethical consideration recorded in point 2 serves to justify the ethical choices made for the AI4REALNET project.

We report this decision in a table similar to (Stefani et al., 2023). Each row corresponds to an ethical issue, and the columns are “Question,” “Decision,” “Consideration,” and “Measure,” respectively. In the first column, the ALTAI question is provided, the second contains the decision decreed either

“Relevant (+)” or “Not Relevant (-)”, the third details the ethical considerations made, and the fourth lists the requirements that are in the UCs responding to the ethical considerations. The resulting tables for the three application domains can be found on the AI4REALNET website<sup>10</sup>.

The summaries of findings in the text form are presented in subsections 2.4.2.1-2.4.2.3 (a more detailed summary is available in Annex 3). Each subsection starts with a figure showing the proportion of questions marked as relevant to the particular ALTAI requirement as an indicator of the ethical dimensions identified as most relevant for the development to be made by AI4REALNET for each UC.

## 2.4.2 SUMMARY OF THE RELEVANT ALTAI REQUIREMENTS

### 2.4.2.1 POWER GRID

The ALTAI assessment of UC in the power grid domain showed the relevance of over 80% for 5 of 7 ALTAI requirements (see Figure 5): accountability, human agency, and oversight, transparency, technical robustness and safety, diversity, non-discrimination, and fairness.

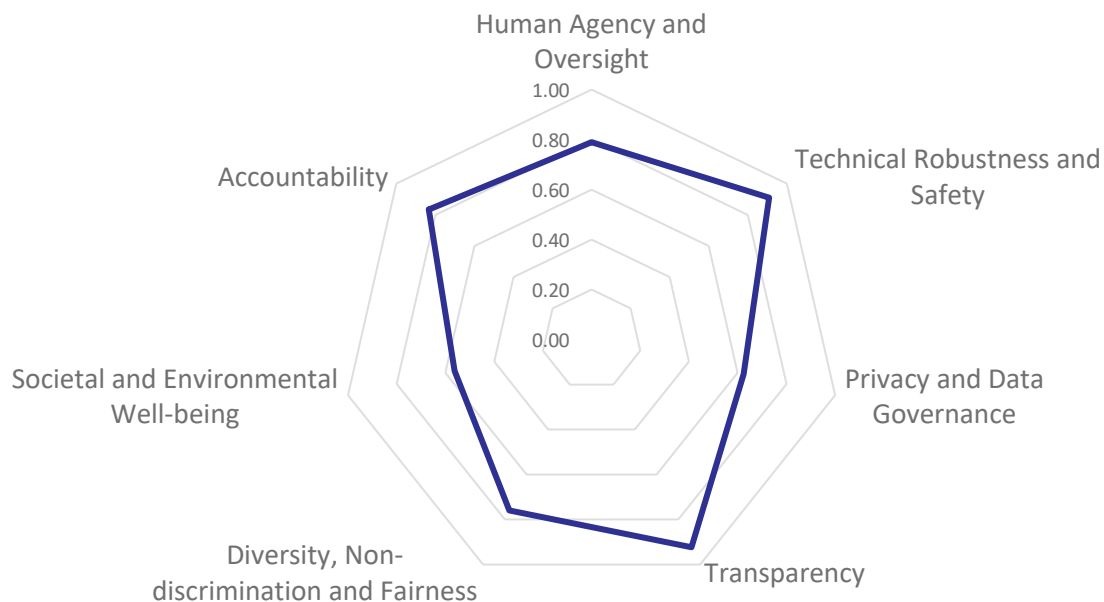


FIGURE 5 – POWER GRID: RELEVANT ALTAI REQUIREMENTS

**REQUIREMENT #1: Human Agency and Oversight.** AI assists human operators in managing power grids by providing recommendations, but human operators retain full control over decision-making. While over-reliance on AI could develop over time, alarms are built into the system when AI cannot make a recommendation, reducing the risk of blind trust. Operators are trained to understand the AI’s reasoning, such as RL, and the system can simulate the impact of recommendations to ensure the operator remains informed and in control.

**REQUIREMENT #2: Technical Robustness and Safety.** AI systems for power grids must be resilient to attacks and data disruptions. Cyberattacks on input data, AI model outputs, and uncertainties in the model are risks. Robustness metrics are necessary to monitor these systems during both training and

<sup>10</sup> [https://ai4realnet.eu/wp-content/uploads/2024/08/D1.1-ALTAI\\_Summary.pdf](https://ai4realnet.eu/wp-content/uploads/2024/08/D1.1-ALTAI_Summary.pdf)



operation. While safety threats like adversarial attacks or environmental risks may arise, AI outputs remain under human control, and inaccuracies will not cause catastrophic outcomes due to human oversight. Transfer learning allows the system to adapt to new environments, and continuous monitoring ensures that AI performance remains optimal. Stress tests will help verify the system's ability to withstand input and model perturbations, ensuring reliability and reproducibility.

**REQUIREMENT #3: Privacy and Data Governance.** The AI system does not handle personal data, so privacy concerns are minimal. Data used for training is anonymized, although operator actions may be traceable through timestamps. The project complies with GDPR requirements, ensuring secure and proper data handling.

**REQUIREMENT #4: Transparency.** Transparency is crucial in AI systems for power grid operations. Transmission operators store historical records of events, ensuring that AI-based decisions can be traced and replayed. While the current AI methods focus on neural networks, feature importance and sensitivity analyses help improve explainability. Communication between the AI system and human operators includes alarms to inform operators of potential AI failures. Training programs are planned to help operators interact effectively with the AI system.

**REQUIREMENT #5: Diversity, Non-discrimination, and Fairness.** The AI system must avoid unfair bias, ensuring it does not favor specific energy producers. Bias may arise from technical limitations of grid operations, but fairness in redispatching or curtailing certain users is essential. Comparing AI decisions with optimal power flow solutions ensures the least-cost outcomes. Stakeholders are involved in the AI design process, and competitions help evaluate the AI's effectiveness and fairness.

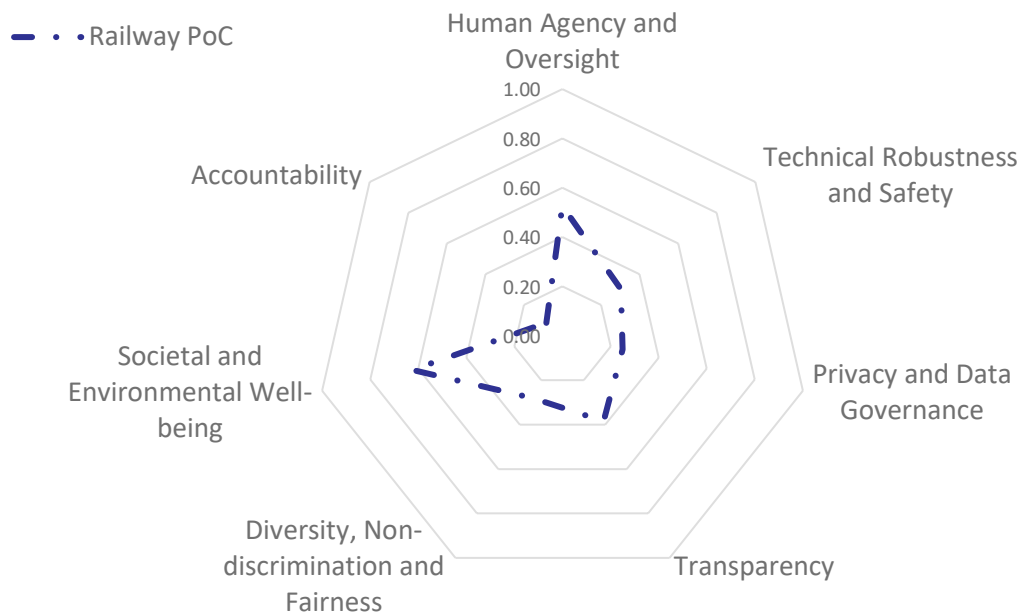
**REQUIREMENT #6: Societal and Environmental Well-being.** AI systems are designed to prioritize carbon-free actions and reduce blackouts. They increase resilience to extreme weather events and aid in minimizing the carbon footprint of grid operations. While AI augments human analytical skills, it does not replace operators. Training programs will enhance operator understanding of AI, ensuring efficient collaboration between human and machine.

**REQUIREMENT #7: Accountability.** Auditability is key for the AI system, especially in cases of outages or cyberattacks. Storing AI model data is crucial for tracing decisions. While audits are unlikely during development, high-risk system regulations, such as the AI Act, will require audits during the operational phase. Risk management systems, including third-party reporting of vulnerabilities, will be necessary to ensure the AI system's safety and reliability.

#### 2.4.2.2 RAILWAY NETWORK

For the railway network, a large proportion of questions in the questionnaire were considered relevant for the UCs but out of scope for the proof of concept (POC) that will be implemented during the AI4REALNET project. The POC is limited to be tested in the simulation environments and is concentrated on the technical feasibility of the functional requirements. Hence, many ethical dimensions will not be included for the first implementation due to the use in a controlled environment but are relevant at later stages. Figure 6 shows the relevant ALTAI requirements and plans for implementation in the AI4REALNET project (Railway PoC). The ALTAI questionnaire on the PoC yields

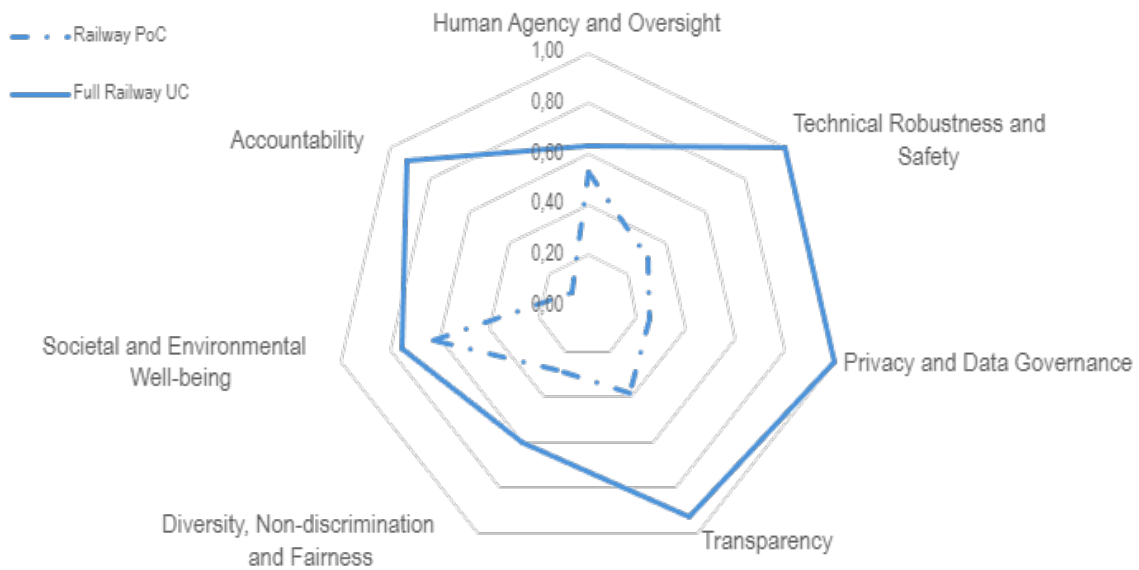
requirements on Human Agency and Oversight, Social and Environmental Well-Being, and Transparency.



**FIGURE 6 – RAILWAY: ALTAI REQUIREMENTS RELEVANT FOR POC PLANNED FOR AI4REALNET**

Additionally, we performed the ALTAI analysis to identify relevant non-functional requirements for the system's future real-world application.

Figure 7 depicts how the number of identified relevant ethical dimensions for the system planned for the application in real-world scenarios increases in comparison to those of the PoC. The dashed line is equivalent to Figure 6 (Railway PoC). The solid line shows the proportion of relevant questions for the real-world applications, to assess the overall coverage of UCs by ALTAI questionnaires. The difference between the dotted line (PoC) and the full line (Full Railway UC) illustrates how some ethical requirements become relevant at later stages of development than the ones covered within the AI4REALNET scope. For the complete coverage of the UCs, requirements such as accountability, technical robustness and safety, privacy and data governance, and transparency have grown in importance. The considerations regarding each of the ALTAI requirements for Railway UCs are summarized below.



**FIGURE 7 – RAILWAY: ALTAI REQUIREMENTS RELEVANT FOR POC AND EXTENDED VERSION FOR THE REAL-LIFE APPLICATION**

**REQUIREMENT #1: Human Agency and Oversight.** The AI system interacts with human end-users, impacting their autonomy and decision-making. Overreliance on the system is a potential risk. Therefore, it is crucial that employees are trained to understand how they are using AI and how to use it properly. While the system doesn't simulate social interaction, it can still foster addictive behavior. In railway operations, human oversight varies from Human-in-the-Loop to Human-in-Command. Procedures must be established to safely revert control back to humans when the AI is the acting agent. Human oversight should include mechanisms for detecting adverse effects and controlling the system's self-learning nature.

**REQUIREMENT #2: Technical Robustness and Safety.** Though the project addresses some aspects of technical robustness and safety, more detailed considerations are necessary when these solutions are implemented. Since collision avoidance is handled separately, the AI system poses minimal risk to human safety. Resilience to attacks is considered, but certification and long-term security procedures fall outside the scope. Safety concerns such as system fault tolerance and technical review require human oversight during development. System accuracy is crucial, with performance monitoring included in the process. Reliability issues should be mitigated with mechanisms to transfer control back to humans and notify them of uncertain AI results. Continuous learning requires documentation and interpretability to ensure system reliability and human control.

**REQUIREMENT #3: Privacy and Data Governance.** The AI system does not use private data, and privacy concerns are minimal in the scope of this project. Future mechanisms to address privacy concerns should be evaluated at later stages of development, although they are not immediately relevant here. Data governance complies with regulations, but GDPR-related measures are not necessary given the nature of the data used.

**REQUIREMENT #4: Transparency.** Traceability is key, enabling historical event records to be replayed, which includes the AI model's input and output data. Explainability ensures human operators understand the AI system's goals, decision-making process, and learning mechanisms. Clear

communication between the AI system and human operators is vital to prevent misuse and build trust. The system is designed to clearly distinguish AI actions, ensuring that human operators are always aware of their interaction with AI and are informed about its capabilities and limitations.

**REQUIREMENT #5: Diversity, Non-discrimination, and Fairness.** Avoiding bias during development is important, ensuring that the AI system fairly distributes delays and does not favor specific Railway Undertaking Operating Managers (RUOMs). Bias detection mechanisms may be developed in the future but are not within the project’s current scope. Stakeholder participation is integral to aligning the system with real-world needs, with workshops involving both stakeholders and the public informing the development process.

**REQUIREMENT #6: Societal and Environmental Well-being.** The AI system could indirectly contribute to environmental well-being through improved efficiency. Its impact on work arrangements and skills is significant, and design considerations must address these changes. Workshops with end-users and human factors experts are recommended to guide development. While the system will require new skills, the creation of training courses is necessary but beyond the current project's scope.

**REQUIREMENT #7: Accountability.** Auditability is ensured through documentation and logging, which are crucial for post-hoc analysis and performance evaluation. Although detailed risk management is not included in the proof-of-concept phase, documentation, and logging provide a foundation for internal AI ethics monitoring and accountability assessments in the future.

### 2.4.2.3 AIR TRAFFIC MANAGEMENT

The summary of the ALTAI questionnaire filled for the ATM UCs is in Figure 8. The requirement of transparency stands out clearly from the others. The focus on its constituents, traceability, explainability, and communication is shaped by the type of AI system described in use cases. For an AI assistant, transparency describes different aspects of the human-AI collaboration and can be used to facilitate the operator's successful use of AI system predictions. The productive cooperation between an operator and an AI system is based on reliable, understandable, and sufficient communication.

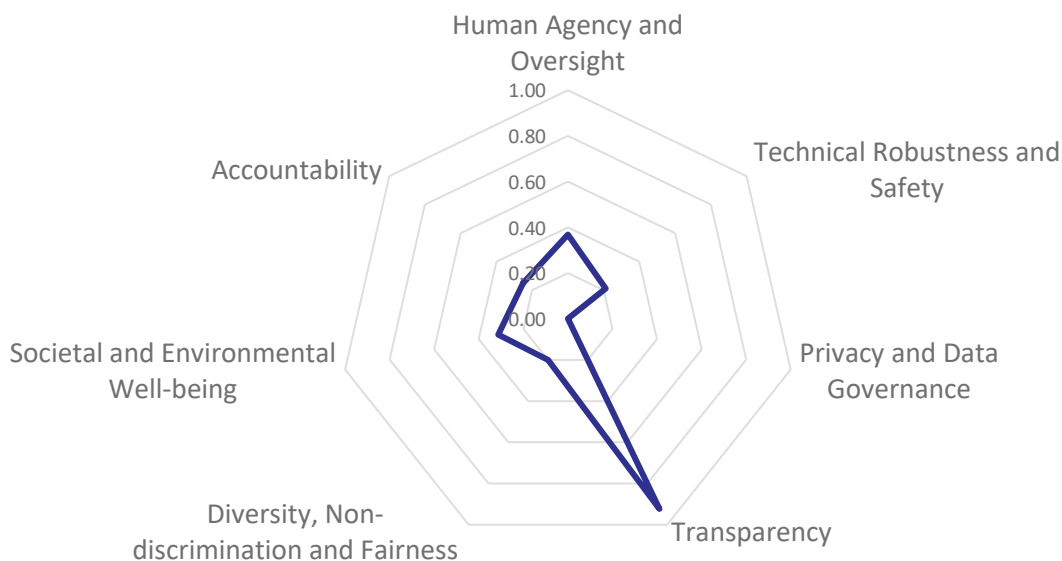


FIGURE 8 – ATM: RELEVANT ALTAI REQUIREMENTS

**REQUIREMENT #1: Human Agency and Oversight.** The AI system operates as a recommender for human operators, and final decisions remain under their control. However, prolonged use of AI might reduce operator vigilance and over-reliance on AI-generated decisions. Currently, there is no risk of addiction or manipulation, but the shift from recommendation-based to fully automated decision-making could affect human autonomy, necessitating stricter rules. As autonomy increases, the operator's oversight decreases, moving toward a "management by exception" model where manual review is minimized. Alarms are triggered if the AI cannot generate a solution or if an environmental change affects the AI's recommendations, ensuring human intervention when necessary.

**REQUIREMENT #2: Technical Robustness and Safety.** At higher automation levels, AI-generated decisions implemented without human confirmation may lead to dangerous situations, making resilience to attacks and system security essential. While the project aims to ensure stability and reliability, risk evaluation must guide the design of safety properties. Any updates to the AI model, particularly with online RL, must be logged and communicated to operators to prevent confusion. Though AI serves as a recommender, low accuracy in suggestions could still lead to adverse outcomes if human oversight falters. Metrics such as KPIs for system accuracy and reliability should be continuously monitored, and fallback plans should be in place, especially when transitioning to automatic implementation of decisions.

**REQUIREMENT #3: Privacy and Data Governance.** No private data will be used during the system's training or operation. However, personal data might be indirectly involved when calculating KPIs, which must be fully anonymized to protect individual identities while preserving the accuracy of performance metrics.

**REQUIREMENT #4: Transparency.** AI system traceability is critical, and all human interventions and decisions should be logged. This includes documenting the input data used to generate decisions to ensure transparency. Explainability is a priority, with operators able to request explanations for AI decisions. Metrics like "Trust in AI solutions" and "Prompt demand rate" will measure operator confidence in AI-generated decisions and the effectiveness of explanations. Regular surveys can be implemented to assess human-system interaction and further improve AI system communication.

**REQUIREMENT #5: Diversity, Non-discrimination, and Fairness.** No specific biases are anticipated in the current system, and the AI primarily serves human operators rather than impacting end-users directly. The introduction of AI into ATC could influence operator workloads, requiring new skills and possibly leading to concerns about job displacement. Stakeholder consultations during the design process will ensure that AI system benefits, risks, and limitations are understood, with feedback gathered through operator surveys for continuous improvement.

**REQUIREMENT #6: Societal and Environmental Well-being.** The AI system is designed to reduce the workload on the air traffic system and decrease carbon emissions, contributing positively to environmental well-being. Metrics to measure carbon savings will be developed. The system also enhances human operator decision-making but requires proper training to mitigate concerns or resistance to changes in work methods. Operators must be educated on the fundamentals of AI to build trust and competence before the system's full implementation.

**REQUIREMENT #7: Accountability.** Auditability is crucial, with AI model weights, hyperparameters, structure, and input data being logged for future verification. Since RL algorithms update continuously, system states should be audited after each update to maintain accountability. The project will make model code publicly available for benchmarking, but operational deployment in real environments would require robust methodologies to ensure traceability and accountability for AI system decisions.

### 2.4.3 RECOMMENDATIONS FOR IMPROVEMENT OF ALTAI

Trustworthiness assessment is more efficient when it addresses issues relevant to each stage of the AI life cycle. ALTAI, being defined mainly as a tool for ex-post analysis, doesn't reflect these nuances. We consider that the utility and efficacy of this tool can be improved by adapting it to allow its application throughout the entire AI Life Cycle. We present the recommendations below for extending ALTAI in this way.

Complementarily, Section 3.3, presents additional analysis of ALTAI with respect to the epistemological and philosophical foundations of trustworthy AI and the concepts of risk and uncertainty. It also provides further suggestions for tailoring ALTAI to critical infrastructure applications.

***1- Create an alternative set of questions, directed to an early development stage that is aimed to highlight checkpoints to consider at respective stages.***

As ALTAI was developed as a post-hoc assessment, the formulation of questions relates to a final product. However, the assessment during the earlier stages is beneficial for introducing trustworthiness requirements at earlier development stages to reach trustworthiness by design and save development time of additional iterations.

It can be compared to the Data Protection Impact Assessment, which is described in the General Data Protection Regulation, Article 35<sup>11</sup>, as “an assessment of the impact of the envisaged processing operations on the protection of personal data”. According to GDPR, this assessment must include a systematic description of processing operations, including the scope and nature of the processing and its functional description; measures applied to comply with regulations; the description of origins, nature, and severity of risks to the rights and freedoms of data subjects as well as “measures envisaged to treat those risks”. GDPR demands DPIA to be conducted “prior to processing” as a means to address any potential risks at an early stage. Additionally, it advises continuing assessment during the entire life cycle.

This approach can also be adopted for the ALTAI: the trustworthiness assessment should be introduced as quickly as possible during the development process to identify and mitigate possible risks and reassess the system after any significant change. It is also advisable to ensure consistency with standards of frameworks for risk identification and management (e.g., *ISO/IEC 23894:2023, Information technology — Artificial intelligence — Guidance on risk management*).

***2- Optimize the assessment process.***

---

<sup>11</sup> Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).

The questionnaire offers a base for algorithm analysis; however, the way it is conducted can have a major influence on the result. The initial suggestion of HELG is to fill the assessment for a particular algorithm. We suggest the following methods to increase the benefits of the ALTAI questionnaire:

Assessment by different groups of stakeholders with subsequent summarization of results. It is possible that depending on the role of the stakeholder, how frequently the user will be facing the algorithm, the role in the organizational processes, the perspective of the algorithm, and expectations from the outcome of its work are different, which shifts the perception of the importance of dimensions of trustworthiness.

Assessment as a part of a co-design workshop. Such a workshop can help to establish the requirements for trustworthiness, engage users and stakeholders in an early development stage, and clarify the vision of the final product. In this case, ALTAI offers a solid base for discussion with a comprehensive list of topics. The Ethics Guidelines for Trustworthy AI also underline that ALTAI can be most valuable in active engagement with its questions.

Regular assessment as a part of the testing procedure. As the algorithms evolve during their development and life cycle, the assessment should be repeated to ensure that the changes introduced in each next version do not sacrifice the established level of trustworthiness. Regular reassessments allow for identifying the changes in the AI system and adjusting the trustworthiness requirements if needed.

***3- Introduce a way to distinguish between irrelevant functionality and those that are currently not in the scope or are not planned.***

If the trustworthiness assessment is performed before the AI system is finalized, it is possible that some functionality mentioned in ALTAI has not been implemented or is not planned for development yet. For example, in AI4Realnet Use Cases from the ATM domain, some security and accountability features are not planned for the prototype but are planned for later implementation. The available answers for the ALTAI questions can be extended to cover these cases.

Due to the specificity of the AI4Realnet project, which contains use cases from 3 different domains, we could observe that assessment in the early stages can be complicated by different understandings of the final product among different groups of stakeholders. In this case, it is hard to generate a robust result if the participants do not agree on the final product. The prerequisite to apply ALTAI at the earlier stages is the comprehensive overview of the future AI system agreed upon among all stakeholder groups.

**4- Find a way to homogenize the number of questions in different dimensions of ALTAI.**

The formulation of ALTAI questions influences the result of the assessment. Currently, the number of questions and subquestions differs a lot among ethical dimensions; some questions have 4-5 subquestions, which are only slightly different from each other. The quantitative analysis of the ethical requirements shown in Figure 5-Figure 8 is influenced by these differences, as one topic can have different contributions to the proportion depending on the overall number of questions in the dimension.

***5- Producing domain/application type-specific versions of ALTAI concentrated around risks relevant to critical infrastructure domains.***

The adapted version should contain only relevant questions and tailor the assessment to the applications and risks that are connected to it. The questions can also be adapted to be applicable at the earlier development stages and be formulated as suggestions and not as a checklist.

The formulation of ALTAI questions has an influence on the assessment results. For example, current coverage of the dimension of Human Accountability and Oversight is directed more to commercial social applications. Some questions are less relevant for the industrial applications covered in the AI4REALNET use cases because of the differences in the UI and the kinds of interactions users have with the AI system. However, such applications are designed for human-machine interaction and need to be assessed accordingly.

Furthermore, if the formulation makes the question not relevant, it means that the coverage of ALTAI is not full and cannot provide a full picture of the state of the system. The functions or properties not mentioned in the ALTAI questions are not evaluated, disregarding their importance for the AI system. This calls for further improvements in the assessment process to increase its suitability for applications in safety-critical infrastructure (see section 3.3).



### 3. CONCEPTUAL FRAMEWORK

A generic overview of the AI4REALNET conceptual framework building blocks is shown in Figure 9, which also summarizes the structure of this section, where each building block is shown with its corresponding subsection. The project followed an interdisciplinary approach to build this framework. The framework combines traditionally separate fields, such as psychology and cognitive engineering, to study how experts make collaborative decisions in complex situations (where automation can have a role) and develop effective design and evaluation criteria for supporting human decision-making. Other fields the framework draws on include mathematics, decision theory, computer science, and specific engineering domains related to energy and mobility. Moreover, for the AI system design, systems engineering and theory adapted to the integration of TAI were applied to construct the operational and functional view and logical architecture of the system to cover the functional and non-functional requirements of the UCs from section 2.

In Section 3.1, the context and decision environment for critical network infrastructures are presented based on the UC scenarios described in the first part of this document. Section 3.2 describes the decision-making process from the **human agent** perspective and as a sociotechnical system (subsection 3.2.1), and the decision-making process from the **AI agent** perspective and the corresponding strategies and methods (subsection 3.2.2). It subsequently details the methodology behind the conceptual framework alongside the system design steps required to conceptualize a generic **human-AI interaction** framework (subsection 3.2.3). Finally, section 3.3 examines the epistemological and normative foundations of the **notion of TAI** and analyses the different components of risk and their application to AI, focusing on critical infrastructures.

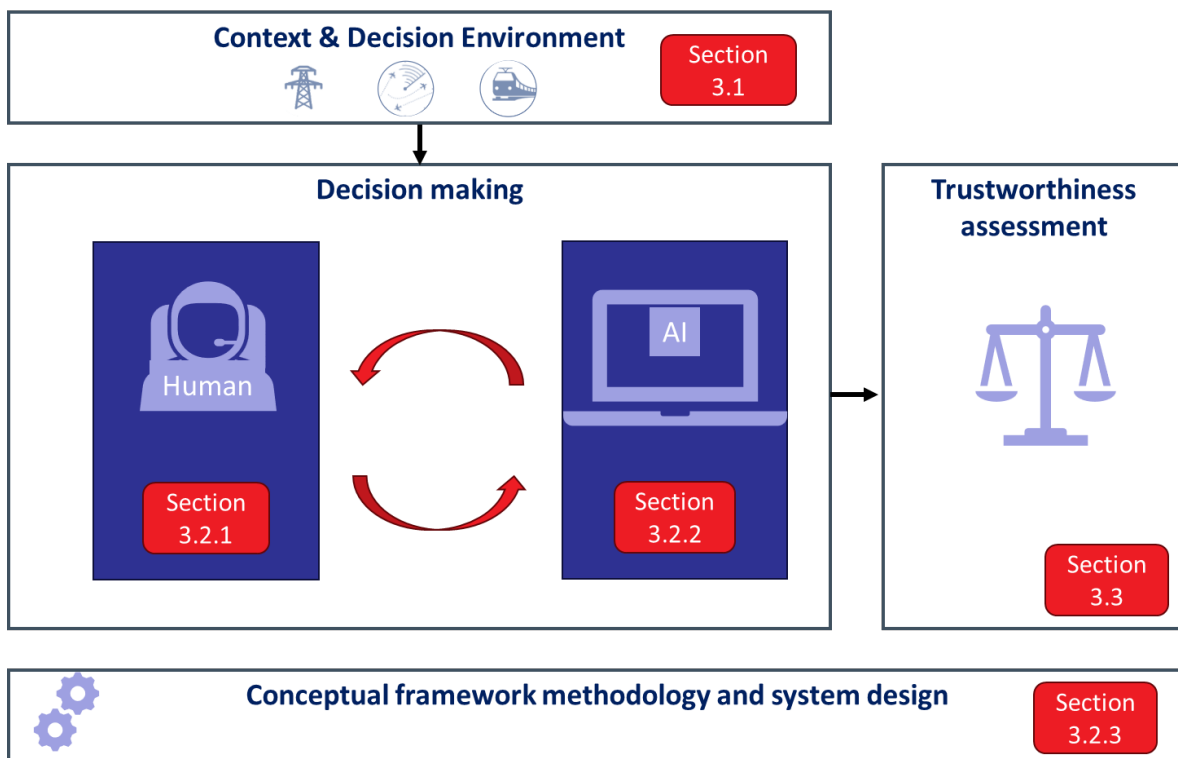


FIGURE 9 – GENERIC VIEW OF CONCEPTUAL FRAMEWORK BUILDING BLOCKS AND SECTION ORGANIZATION

### 3.1 CONTEXT AND DECISION ENVIRONMENT

#### 3.1.1 DECISIONS ON CRITICAL NETWORK INFRASTRUCTURES

Decisions in critical network infrastructure operations are at the heart of operational processes in critical network infrastructures. They can be described in three main points (see Figure 10 and Table 6). Firstly, they are made to manage constraints on a network capacity that can stem from external events (operational disruptions or emergencies) and are detected through observations or forecasts of the infrastructure’s state that include a certain level of uncertainty and external context. Secondly, they also involve multiple operators or stakeholders from short to long-term horizons. Lastly, they are made under time constraints and trade-offs between multiple and conflicting objectives and lead to both preventive and corrective actions that are chosen within a large action space and are planned or implemented in real-time, respectively.

Examples of decision-making scenarios are given in Annex 4, which are about context, characteristics, impacts, and evaluation of decisions.

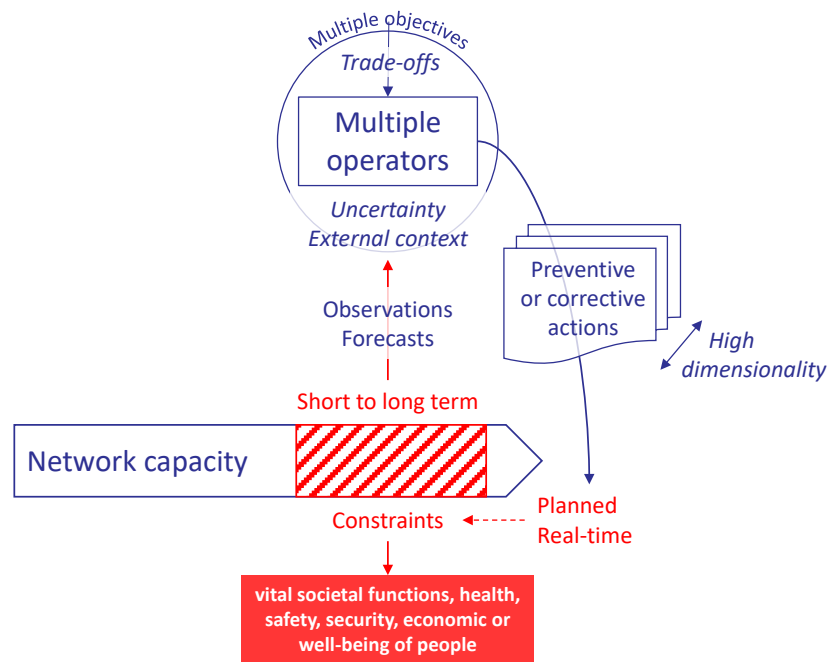


FIGURE 10 – DECISIONS IN CRITICAL NETWORK INFRASTRUCTURE OPERATIONS

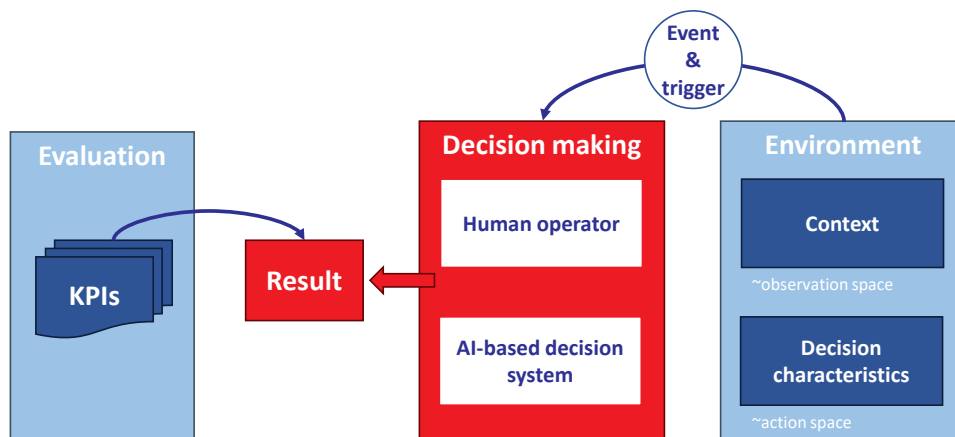
The criticality of the decisions is directly linked to the critical nature of the underlying infrastructure for ensuring vital societal functions, health, safety, security, economic or well-being of people, namely:

*“European critical infrastructure means an asset, system or part thereof located on EU territory, which is essential for the maintenance of vital societal functions, health, safety, security, economic or well-being of people, and the disruption or destruction of which would have a significant impact on at least two Member States, as result of the failure to maintain those functions. The significance of the impact*

*is assessed against distinct cross-cutting criteria, which encompass casualties, economic and environmental effects and public effects.”<sup>12</sup>*

The decisions on critical network infrastructures can be analyzed based on the following framework (see Figure 11), which is centered on the decision-making and includes:

- Prerequisites to make a decision, that is, the environment in which the decision is made, composed of a context<sup>13</sup> (e.g., network infrastructure, events) and characteristics<sup>14</sup> of a decision,
- Consequences, or impacts of a decision, that is its results,
- Evaluation of a decision.



**FIGURE 11 – DECISIONS ANALYSIS OF CRITICAL NETWORK INFRASTRUCTURES**

The decision-making step itself is triggered by a given event detected in the environment. It involves both the human operator and the AI-based decision system, who interact in multiple ways (manual, co-learning, and autonomous). It is composed of back-and-forth iterations between exploration and validation/feedback tasks, as depicted in Figure 12.

<sup>12</sup> Source: Directive 2008/114/EC, Articles 2 and 3

<sup>13</sup> In a Reinforcement Learning context, this can be referred to as the “observation space”

<sup>14</sup> In a Reinforcement Learning context, this can be referred to as the “action space”

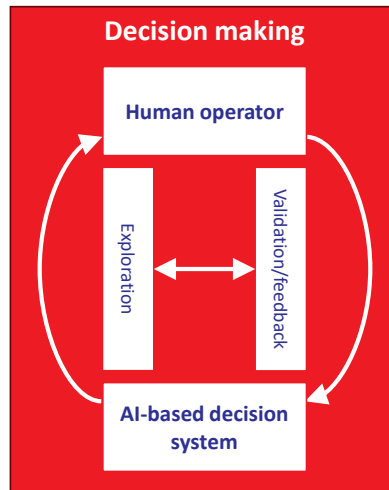


FIGURE 12 – DETAIL OF DECISION MAKING

Common works have facilitated the identification of these common steps across all project domains centered around illustrative examples of various operating scenarios (described more in detail in Annex 4).

### 3.1.2 CONTEXT, CHARACTERISTICS, IMPACTS, AND EVALUATION OF DECISIONS

To extract the common aspects of decisions across the three types of critical infrastructures studied (see section 3.1.1). Thus, to provide a better description of the decision process from a business perspective, an analysis was performed on data collected using a detailed questionnaire for each domain (the data can be found in Annex 4). This questionnaire is structured into four main topics: context, characteristics, impacts, and evaluation of decisions.

Based on all data collected, a similarity score has been performed across pairs of domains to give an idea of how much similarity exists across the three domains (the methodology is detailed in Annex 4):

Decision analysis	Air Traffic-Electricity	Electricity-Railway	Railway-Air Traffic
Context	13%	13%	12%
Characteristics	40%	50%	50%
Impacts	0%	4%	0%
Evaluation (KPIs)	23%	38%	46%

TABLE 5 – SIMILARITY SCORE OF DECISION ANALYSIS ACROSS DOMAINS

Even if the decision context is different for each domain (which can be explained by the fact that each domain remains intrinsically different), we observe that the characteristics of the decision have a higher degree of similarity, which is of the same level of magnitude across the different pairs of domains. This illustrates the interest in performing multi-domain work.

The following table lists the words describing the decision process that are similar across all three domains, for decision context and characteristics:

Category	Similar words
<b>context of the decision process</b>	<ul style="list-style-type: none"> <li>external events,</li> <li>multiple operators</li> </ul>
<b>decision characteristics</b>	<ul style="list-style-type: none"> <li>preventive or corrective,</li> <li>planned or real-time,</li> <li>large and mixed action space,</li> <li>real-time to long-term.</li> </ul>

**TABLE 6 – SIMILAR DECISION CHARACTERISTICS ACROSS ALL DOMAINS**

Then, the second highest similarity is obtained on the evaluation part, with the following KPIs similar across all domains:

- assistant relevance,
- trust in the AI system.

Within the multi-domain work, this shows the interest of the evaluation that will be carried out in WP4.

On the other hand, the level of similarity across domains is almost zero for the “impact of a decision” topic: this can be explained by the very domain-specific impacts of each decision. In line with the similarity scores obtained for the “impact of a decision” topic, there are no similar words across all domains for this topic.

Finally, we can observe that the two most similar pairs of domains are “Railway-Air Traffic” (highest) and “Electricity-Railway”.

## 3.2 DECISION-MAKING PROCESS

### 3.2.1 HUMAN AGENT AND DECISION-MAKING

This section describes the AI4REALNET framework from an overarching perspective and from a sociotechnical systems perspective. The main assumption is that all work systems are sociotechnical as they follow different principles, which must be considered when combining them. Only joint optimization increases the performance of the work system as a whole. In contrast, optimization of one sub-system may decrease overall performance. This is because the two sub-systems interact and hence may empower or depower each other.

Regarding AI4REALNET, this leads to two main conclusions. First, AI design needs to take requirements derived from characteristics of the social sub-system (i.e., human factors) into account. Second, to be able to exploit AI capabilities and potentials, the social sub-system must also be designed accordingly. This refers to social aspects such as human skills, process design, or even organizational culture. If, for example, AI is designed as a system providing recommendations, it is the human’s role to judge these recommendations and to decide. This requires, but is not sufficient for, appropriate skills. In addition, a corresponding task design is required. Finally, the leadership style also needs to fit the concept of the envisioned role. This is because humans can make mistakes and, therefore, wrongly reject a recommendation generated by AI. A likely consequence of being blamed for this mistake is that the person concerned will no longer have the confidence to reject AI-based recommendations and will

blindly nod them off. This cultural effect contradicts the original AI design, which depends on an engaged human decision-maker.

General principles for sociotechnical system design are well elaborated, e.g., (Clegg, 2000). However, with regard to AI integration into sociotechnical systems, there is still significant research required, as stated, e.g., by the National Academies of Sciences, Engineering, and Medicine (2020), by (Endsley, 2023a), or by (Naikar et al., 2023).

Against this background, three aspects of AI integration in sociotechnical systems relevant to AI4REALNET are examined in the following sections, namely different design approaches and their effect on human behavior, normative aspects of AI design, and descriptive aspects of AI design.

### 3.2.1.1 DIFFERENT DESIGN APPROACHES AND THEIR EFFECT ON HUMAN BEHAVIOR

When introducing Information Technology into work processes, two fundamentally different strategies can be pursued: “automate” versus “informate” (Zuboff, 1988). While automation aims at replacing human skills and human effort with technology, information aims at complementing humans with technology. In practice, full automation has not (yet) been achieved for complex work processes; there are always humans involved when supposedly autonomous systems are integrated into sociotechnical systems (Bradshaw et al., 2013). However, if only those functions are allocated to humans that cannot be automated for technical or cost reasons, the result is an unaccomplishable task for humans. Bainbridge (Bainbridge, 1983) described this left-over approach as “Ironies of Automation”. The main problem is that humans need to monitor technical performance, which causes problems like monotony and fatigue. Beyond this, humans may lack the capabilities required to supervise a technology that was designed to act faster and take more factors into account than humans are able to. Other negative effects of the left-over approach include, for example, over-confidence and under-confidence in technology, as well as misjudgment of process states, inadequate situation awareness, demotivation, or loss of skills and experiences as a result of automation (e.g., Manzey, 2012).

Avoiding such negative effects on the human contribution to system safety and reliability is particularly important for AI4REALNET. This is because AI4REALNET aims to develop AI support to improve the resilience of critical network infrastructures. ATM is one of the networks focused on by AI4REALNET. In a white paper on resilience engineering, EUROCONTROL – the European Organisation for the Safety of Air Navigation – states ANSPs are increasingly confronted with instability and variability and that “this requires them to be flexible, to rely on human ingenuity and skill (...)” (EUROCONTROL, 2009, p.8). The same paper defines resilience as the “(...) intrinsic ability of a system to adjust its functioning prior to, during, or following changes and disturbances so that it can sustain required operations under both expected and unexpected conditions” (EUROCONTROL, 2009, p.2). It is common sense in the resilience engineering community that humans crucially provide system resilience (e.g., Hollnagel et al., 2006). With regard to AI, (Naikar et al., 2023) state that “The features of emerging AI technologies, assessed together with the properties of complex environments, suggest that their relationships to humans may need to become increasingly collaborative in nature.” (p. 1688).

To enable human-technology collaboration, technology needs to be designed and implemented into organizational processes in a way that takes human characteristics into account (e.g., Grote et al., 1995). For AI, the corresponding design requirements have been described in (Wäfler and Rack, 2021;

Endsley, 2023b; Miller, 2023). These requirements take into account the fact that humans and technology are qualitatively different, even though technology, due to machine learning (ML), has developed impressive capabilities. Humans are distinguished by their understanding, commitment, and ability to take responsibility. To fully activate and sustain these traits, tasks should be designed to keep them consistently engaged. For example, humans must have an active role in task performance in order to maintain situational awareness. If a passive role only is assigned to humans, their attention and ability to concentrate will be impaired. A certain degree of autonomy and self-determination is a prerequisite for humans to be in a state of interest and commitment to the task (i.e., intrinsic motivation). If prerequisites like these are not met, humans will not be able to contribute their potential to the joint human-AI performance. Various factors affect these prerequisites. One of them is AI design. It can support an active role of the human or impose a passive role. It can increase human autonomy or take control of the human. In the following sections, corresponding normative design requirements will be described for four human-oriented objectives: human decision-making, human motivation, human learning, and human trust in AI.

### 3.2.1.2 NORMATIVE ASPECTS

Miller (2023) describes five types of how AI can support decision-making in human-AI collaborative systems, i.e., five types of explainable AI (XAI):

- Recommendations without explanations: AI provides suggestions for decisions without any further explanations.
- Recommendations with explanations: AI provides suggestions for decisions with further explanations.
- Recommendations with interpretable model: AI makes its decision model transparent.
- Cognitive forcing: The Human makes the initial decision; AI provides explanations and recommendations regarding this human-initialized decision.
- Evaluative AI: The Human formulates a hypothesis, and AI provides the human with evidence for and against this hypothesis.

The following section describes how these five types of XAI help or hinder human decision-making, human motivation, human learning, and human trust in AI. Against this background, the consequences of the three AI4REALNET scenarios: i) AI-assistant to human (human in control), ii) joint human-AI decision-making (including human-AI co-learning), and iii) autonomous AI (human as a supervisor) are reflected.

### 3.2.1.3 HUMAN DECISION-MAKING

Today's AI-based decision-support systems are mainly based on recommendations. However, recommendations provided by AI are usually not sufficient, even if they are enriched by means of explanations (XAI) and transparency (Eisbach et al., 2023; Miller, 2023). Several studies showed that explanations do not automatically lead to better decisions (Ngo & Krämer, 2022; Zhang et al.). Therefore, rather than just providing decisions, joint human-AI decision-making based on the complementary capabilities of humans and AI is required (Endsley, 2023; Miller, 2023). From a psychological perspective, joint decision-making needs to consider the human decision-making processes with its cognitive elements as well as with its related biases such as the anchoring effect or the confirmation bias (Eisbach et al., 2023; Ha & Kim, 2023; Wang et al., 2019).

Human decision-making goes beyond mere choice between options. Rather, it is a multifaceted cognitive process that aims to make sense of and understand complex and dynamic environments in order to make meaningful decisions (Endsley, 2023b; Hoffman et al., 2009; Klein et al., 2003; Klein, 2018). Macrocognitions such as problem detection, attention management, and anticipation are key in this process. Furthermore, effective decision-making depends on profound operational knowledge, enabling further macrocognitions such as process monitoring and situation awareness, allowing for timely intervention when needed.

The design and deployment of AI are changing human tasks and, consequently also, the conditions for carrying out these cognitive processes. It has an impact on human behavior and perception and, ultimately, on decision outcomes (Endsley, 2023b; Parker and Grote, 2022). AI must, therefore, be designed and deployed in a way that supports these cognitive processes to ensure the quality of decisions.

In the following sections, relevant cognitive processes related to 1) developing a thorough knowledge of the operational process, 2) enhancing process monitoring, 3) achieving a comprehensive understanding of the current situation, and 4) mitigating cognitive biases are described.

#### **3.2.1.3.1 KNOWLEDGE OF THE OPERATIONAL PROCESS**

At its core, knowledge of the operational process refers to the critical infrastructure that needs to be controlled. It includes but is not limited to knowledge about system behavior in terms of knowing how the system operates, how it behaves under normal and non-normal conditions, and how it responds to various inputs from its environment. This also includes knowledge of leverage points for influencing processes. This expertise is represented in mental models (Endsley, 2000; Klein, 2018; Klein et al., 2003). Consequently, AI must provide insights that are suitable for developing, maintaining, and refining the corresponding mental models of human decision-makers.

#### **3.2.1.3.2 MONITORING THE OPERATIONAL PROCESS**

Monitoring the operational process through AI is central to understanding the real-time process status and making informed decisions based on current conditions. This requires knowing what to look for and constant attention management (Endsley, 2023a; Klein, 2018; Klein et al., 2003). In this way, AI must reveal both potential problem areas in the operational processes and support humans in developing adequate mental models.

#### **3.2.1.3.3 UNDERSTANDING THE CURRENT SITUATION**

Detecting and comprehending problems, as well as anticipating further developments, serves to understand the current situation, resulting in situation awareness (Endsley, 2000, 2023b, 2023a; Klein, 2018; Klein et al., 2003). This is associated with knowing what to expect from the future situation and knowing what to do. In this way, AI must support humans in the continuous development, updating, and refinement of situational awareness in relation to the current operation of the system and possible future states.

#### **3.2.1.3.4 MITIGATING COGNITIVE BIASES**

The anchoring effect describes a remarkably robust cognitive bias that influences human judgment and decision-making (Furnham & Boo, 2011; Pohl, 2006). It describes the phenomenon that initially presented information “anchors” people’s attention and perception, making them blind to other information (Tversky & Kahneman, 1974). It emerges regardless of different factors such as motivation,



cognitive load, expertise, or even types of anchors (Furnham & Boo, 2011). However, it is known that “the higher the ambiguity, the lower the familiarity, relevance or personal involvement with the problem, a more trustworthy source or plausible bid/estimate, the stronger the anchoring effects” (Furnham & Boo, 2011, p. 37). The anchoring effect is characterized by its long-lasting effects (Pohl, 2006; Wilson et al., 1996). Even an explicit communication of the anchoring effect does not mitigate the effect (Wilson et al., 1996).

Confirmation bias refers to the tendency to seek confirmation of one’s own assumptions by selectively searching for, interpreting, and remembering information in a way that systematically hinders the possibility of rejecting one’s own assumptions (Pohl, 2006). Confirmation bias, therefore, leads to information that contradicts one’s own assumptions being neglected, which causes distorted decisions.

To overcome the confirmation bias in human-AI collaboration, Ha and Kim (Ha and Kim, 2023) suggest providing the human with a priori information (e.g., a set of data that is taken into account when computing decisions) before showing the final decisions generated by AI. According to these authors, this might be the only way to effectively overcome the confirmation bias. In contrast, there are still no ways to fully overcome the anchoring effect when AI suggests recommendations (Pohl, 2006; Wilson et al., 1996; Furnham & Boo, 2011).

**3.2.1.3.5 EFFECTS OF AI APPROACHES ON HUMAN DECISION-MAKING**

Table 7 evaluates what impact different XAI approaches (according to Miller, 2023) have on human decision-making (i.e., on macrocognition and cognitive biases).

Type of XAI (Miller, 2023)	Macrocognition	Cognitive biases
<b>Recommendations without explanations</b>	-/-	-/-
<b>Recommendations with explanations</b>	+/-	-/-
<b>Recommendations with an interpretable model</b>	+/-	-/-
<b>Cognitive forcing</b>	+/-	+/-
<b>Evaluative AI</b>	+/+	+/+

*Note. Each type of AI (Miller, 2023) is evaluated to determine the extent to which it supports the macrocognitive functions and processes and counteracts the cognitive biases. The scoring is as follows: -/- = no support/low counteraction; +/- = partial support/medium counteraction; +/+ = fully supported/high counteraction.*

**TABLE 7 – TYPE OF XAI RELATED TO MACROCOGNITION AND COGNITIVE BIASES**

Human decision-making is a complex cognitive endeavor involving many macrocognitive processes and functions in the human’s brain (e.g., detecting problems, managing attention, sensemaking, and maintaining situation awareness). An AI supporting human decision-making needs to explicitly support these processes and functions. An AI that simply makes recommendations does not support such human cognitive process and hence does not support the human decision-making process.

The evaluation of different types of XAI (according to Miller, 2023) with respect to their support of macrocognition indicates that recommendation-based approaches, even with explanations and interpretable models, fail to foster adequate situation awareness due to a lack of active human involvement in decision-making. Even if AI provides sophisticated explanations or interpretability, humans are not aware of the situation before the AI's recommendations. This has implications for all macrocognitive functions and processes, as they are highly interdependent. As a result, quick and appropriate human evaluation of AI-generated decisions to deal with sudden events is most likely not possible.

When there is sufficient time for humans to evaluate AI-generated decisions, recommendations without explanations still provide insufficient support for macrocognition, as they lack transparent reasoning and are, therefore, not comprehensible. In contrast, XAI approaches that provide explanations or interpretable models may offer partial support for some macrocognitive functions and processes, but only if explicitly designed for that purpose. However, for full macrocognitive support, sophisticated and multifaceted explanations are required (e.g., by evaluative AI). For example, to support the macrocognitive function of "detecting problems", explanations must clarify the reasons for detected problems by identifying contributing factors or patterns. Similarly, the environmental changes that affect decisions should be detailed in explanations that support the macrocognitive function "adapting".

In addition, recommendations, regardless of explanations or interpretable models, can trigger anchoring effects and confirmation bias, likely leading to inappropriate evaluation of AI-generated decisions.

Enhanced XAI approaches, such as cognitive forcing and evaluative AI, involve humans in the decision-making process and are, therefore, better suited to support macrocognitive functions and processes. The involvement and support of macrocognition are greater with evaluative AI than with cognitive forcing. Cognitive forcing allows the human to initiate decisions (i.e., involving the human in setting the topic), but AI still only provides explanations and recommendations. In contrast, evaluative AI allows the human to formulate a hypothesis (i.e., involving the human in reasoning) while AI provides evidence for and against the human-generated hypothesis (supporting the human decision-making process).

Evaluative AI goes beyond simply providing recommendations and explanations. It supports, among other things, sensemaking and maintaining situation awareness, as humans are cognitively involved in all phases of decision-making at any time. This is the prerequisite for the ability to respond quickly and appropriately to sudden events. Evaluative AI is also more effective at mitigating the confirmation bias by actively assisting humans in generating their own solutions, and it not only provides evidence supporting the human's assumptions, but also provides evidence against them. Furthermore, the latter supports identifying new or erroneous patterns, thereby facilitating the refinement of mental models and other macrocognitive functions and processes. Nevertheless, addressing the anchoring effect remains a significant challenge.

#### **3.2.1.3.6 GENERAL CONCLUSIONS REGARDING AI4REALNET SCENARIOS**

In the AI4REALNET project, different scenarios will be developed and implemented, namely *AI-assistant to human (human in control)*, *joint human-AI decision-making (including human-AI co-*

*learning*), and *autonomous AI (human as a supervisor)*, which differ in their consequences for the decision-making process.

Recommendation-based AI, as envisioned in the first scenario of AI-assistant to human (i.e., human in control), does not fully contribute to supporting macrocognition effectively. Although approaches with explanations or interpretable models may partially help improve understanding of the underlying decision-making basis, they do not meet many of the requirements sufficiently. Complete support for macrocognition requires complex and multifaceted explanations and active involvement of humans in the decision-making process. In addition, recommendations can trigger and reinforce anchoring effects and confirmation biases. Recommendation-based AI fails to help humans overcome these biases, as well as their own assumptions and misconceptions. The primary reason for these limitations is that recommendation-based AI solely provides recommendations without supporting humans in their own decision-making process and does not address the biases of either humans or AI.

The scenario of joint human-AI decision-making (including human-AI co-learning) represents a significant advancement regarding the support of human decision-making processes due to several factors. Firstly, the active involvement of humans throughout the decision-making process supports sensemaking and other macrocognitive processes of human decision-making. This not only leads to a better understanding of how decisions are made but also to more informed decisions. It also helps humans to maintain situational awareness so that they can react quickly in urgent situations. Secondly, AI may also help humans identify new patterns and evaluate existing assumptions by providing evidence for and against these assumptions. This supports the mitigation of confirmation bias. Nevertheless, addressing the anchoring effect, which can even be triggered by the reaction of the AI, remains a significant challenge.

The third scenario, autonomous AI (human as a supervisor), does not involve humans in decision-making and, therefore, cannot be assigned to any of the XAI types described by Miller (Miller, 2023). It poses significant challenges regarding macrocognition and overcoming cognitive biases. It does not support any macrocognitive functions and processes, nor does it help humans overcome cognitive biases. This is due to the reduction of the human role in monitoring the AI, resulting in low situational awareness and, therefore, a higher probability of inappropriate decisions when the situation requires human intervention.

#### **3.2.1.4 HUMAN MOTIVATION**

The tendency to not use IT tools (Fildes et al., 2009) and algorithm aversion is quite common (Niehaus et al., 2022; Schaap et al., 2023). Therefore, intrinsic motivation to use AI must be deliberately promoted. Intrinsic motivation is triggered by task orientation - i.e., the state of a human's interest in and commitment to a task (Hackman & Oldham, 1976; Parker & Grote, 2022). Consequently, task design has an impact on a human's motivation to perform and to achieve task-related objectives.

Any automation that supports task fulfillment changes the contribution required from the human and thus the human's task. Since AI automates at least parts of the task, the design of the AI and the way AI is used have a direct influence on human motivation.

Key task aspects that influence human motivation are meaningfulness, autonomy, and feedback (Hackman & Oldham, 1976; Morgeson et al., 2005; Parker & Grote, 2022). AI and the way AI is used

must be specifically designed to have a positive impact on these aspects. Corresponding challenges and requirements for the design and use of AI are described below.

#### 3.2.1.4.1 MEANINGFULNESS

Task-related meaningfulness means that humans experience meaningfulness in what they do (Hackman & Oldham, 1976; Parker & Grote, 2022; Sadeghian & Hassenzahl, 2022). This means that AI must provide them with answers to the question of why they do what they do.

#### 3.2.1.4.2 AUTONOMY

Task-related autonomy means that humans are provided with options between which they can choose (Hackman & Oldham, 1976; Morgeson et al., 2005; Schaap et al., 2023). However, pseudo-autonomy must be avoided. For example, the choice between using an AI or not using it is considered pseudo-autonomy. Rather, the possibility to choose between different ways of using AI offers a real choice. Similarly, the mere acceptance or rejection of AI-generated suggestions is considered pseudo-autonomy. Instead, the choice between several possible solutions is considered to provide autonomy to the human.

#### 3.2.1.4.3 FEEDBACK

To be motivated, people need feedback on their work. If people do not know (do not receive feedback) whether they have achieved their goals or not, they lose motivation. At its core, task-related feedback has two purposes. On the one hand, humans need to know how well they have achieved the objectives of their tasks once they have completed them. On the other hand, humans need to know whether they are on the right track when they fulfill the task. Both types of feedback should be provided promptly (Hackman & Oldham, 1976; Parker & Grote, 2022). Such feedback can be supported, for example, by an AI analyzing the effects of decisions made by humans or showing humans what effect decisions other than those made would have had.

#### 3.2.1.4.4 EFFECTS OF AI APPROACHES ON HUMAN MOTIVATION

The following example illustrates how different AI approaches, according to (Miller, 2023), are rated regarding the three key task aspects that influence human motivation (Hackman & Oldham, 1976). The scores are also shown in Table 8. The table shows an assessment of the extent to which the prerequisites for motivation (experienced meaningfulness, experience responsibility, and knowledge of results of own work as described in the three sections above) are supported by the different types of XAI described by Miller (2023).

In the medical field, a physician relies on an AI system to analyze x-ray images and make diagnoses. If the AI provides recommendations without any explanations, the physician's sense of meaningfulness, responsibility, and knowledge of their own effectiveness diminishes due to the lack of traceability, autonomy, and feedback. The lack of transparency makes the physician feel disconnected from the decision and the decision-making process, resulting in reduced intrinsic motivation. This, along with their inability to explain to the patient why the AI-recommended therapy was chosen, likely results in the recommendation being disregarded.

Conversely, when the AI provides explained recommendations or interpretable models, decision transparency, as well as model transparency, increases. This allows the physician to at least partially understand the AI's reasoning, leading to a clearer understanding of the diagnosis and the diagnostic process, enabling them to explain it to the patient. However, despite this transparency, the doctor's

involvement in decision-making remains limited, which prevents full cognitive engagement and leads to a decreased sense of meaningfulness and responsibility. Also, no transparency is provided regarding the effectiveness of human decisions. This transparency is important for motivation. Suitable AI can provide such feedback: The physician can accept, reject, or modify the AI's suggestion. However, they do not receive any explicit feedback on whether they have made the right decision or whether a different decision would have been better. AI could help here by also showing the physician the effects of their decision. With an evaluative AI, for example, the physician could formulate their expectations regarding the effect of their decision as a hypothesis, and the AI would provide evidence pro and contra their expectations. This would give the physician explicit feedback on their decisions and thus have a positive effect on their motivation.

Alternatively, human-centered approaches such as cognitive forcing and evaluative AI empower the physician to take the lead in decision-making. With cognitive forcing, the physician initiates the decision-making process, which not only increases their control but also forces them to reflect on the decision-making process. While the former increases their sense of responsibility, the latter results in a clearer understanding of the “why” and hence provides a sense of meaningfulness. Both foster intrinsic motivation and engagement with the AI system.

Evaluative AI has even greater potential than cognitive forcing to foster intrinsic motivation, as the physician formulates hypotheses about diagnostic and treatment options for which the AI provides evidence in favor and against. The physician is even more involved in the decision-making process, which reinforces their sense of purpose and responsibility. If evaluative AI also supports the assessment of whether diagnosis and treatment options have led to the expected effects, this will also support the physician's sense of their own effectiveness. Consequently, evaluative AI has the greatest potential to foster intrinsic motivation.

Type of XAI (Miller, 2023)	Experienced meaningfulness	Experienced responsibility	Knowledge of results of work
<b>Recommendations without explanations</b>	-/-	-/-	-/-
<b>Recommendations with explanations</b>	+/-	-/-	+/-
<b>Recommendations with an interpretable model</b>	+/-	-/-	+/-
<b>Cognitive forcing</b>	+/+	+/+	+/-
<b>Evaluative AI</b>	+/+	+/+	+/+

Note. Each type of AI (Miller, 2023) is evaluated to determine the extent to which it supports the corresponding critical psychological state according to the Job Characteristics Model (Hackman & Oldham, 1976). The scoring is as follows: -/- = no support; +/- = partial support; +/+ = fully supported.

**TABLE 8 – TYPE OF XAI RELATED TO THE CRITICAL PSYCHOLOGICAL STATES AND THEIR EXPRESSION**

**3.2.1.4.5 GENERAL CONCLUSIONS REGARDING AI4REALNET SCENARIOS**

In the AI4REALNET project, different scenarios are considered, namely *AI-assistant to human*, *joint human-AI decision-making (including human-AI co-learning)*, and *autonomous AI*, are developed and

implemented, which differ in their consequences on the development and maintenance of intrinsic work motivation.

The first scenario, AI-assistant to human (human in control), is a recommendation-based decision-support that can come with or without explanations or an interpretable model. All variants have in common that they do not involve the human in decision-making but assign him the role of assessing a recommendation. This poses several challenges for intrinsic work motivation and humans' ability to take on the role assigned to them. These challenges include 1) a lack of deep understanding of the decision and the decision-making process, reducing the sense of meaningfulness, 2) limited autonomy if the AI provides only one recommendation, leading to a lack of sense of responsibility, and 3) no feedback on the decision's effect resulting in low knowledge of results of own work. This occurs as the AI initiates and is solely involved in recommendation generation, while the human only selects among recommendations without prior involvement. However, explanations may provide some understanding and contribute to experiencing meaningfulness. Including feedback on decisions' effects in explanations enhances knowledge of the results of work. Generally, relying on recommendation-based AI without human involvement, as in this scenario, leads to low intrinsic work motivation and creates barriers to high performance in human-AI collaboration. To compensate for low experienced meaningfulness and low experienced responsibility, humans may choose not to use AI support and instead make decisions themselves.

Scenario two, which is joint human-AI decision-making (including human-AI co-learning), involves the human in the decision-making process. This can be realized by cognitive forcing and evaluative AI. Both are promising approaches to human-AI collaboration regarding intrinsic work motivation. By involving humans throughout the decision-making process, experienced meaningfulness is positively influenced. In addition, the active involvement of humans in the decision-making process gives them a degree of control over what the AI is doing, which increases their sense of autonomy. This allows the human to initiate recommendation (cognitive forcing) or even formulate a hypothesis (evaluative AI), resulting in experienced responsibility for the results of the work. In addition, evaluative AI may provide humans with comprehensible feedback regarding the effectiveness of their decisions and, therefore, increases intrinsic motivation to participate in the collaboration actively.

The third scenario of autonomous AI (human as a supervisor) does not involve humans in decision-making and, therefore, cannot be assigned to any of the XAI types described by Miller (Miller, 2023). It poses significant challenges regarding intrinsic motivation as it neither supports experienced meaningfulness nor experienced responsibility or knowledge of the results of one's own work. This approach has no benefits regarding intrinsic work motivation. In addition, the implementation of fully autonomous AI results in humans only taking on monitoring tasks, which relates to the irony of automation (Bainbridge, 1983): humans will lack situation awareness but are expected to intervene when necessary, leading to potential breakdowns in effectiveness and increased risks in task execution.

### 3.2.1.5 HUMAN LEARNING

Human learning is a multifaceted process that incorporates psychological, physical, and social dimensions, shaping our perception and interaction with the world (Alexander et al., 2009). At its core, the experiential learning theory proposed by David Kolb, (1984) outlines a cyclical four-stage model—*concrete experience, reflective observation, abstract conceptualization, and active experimentation*.

This model emphasizes the dynamic and iterative nature of learning, where individuals continuously engage with experiences to acquire and refine knowledge (A. Y. Kolb & Kolb, 2009). In the context of decision-making, a deliberately designed learning process is crucial for the human decision-maker to develop a thorough understanding of the subject matter of decision-making (i.e., the process), the decision-support tool (i.e., the AI tool), and the human decision-makers (i.e., learning about oneself). The following sub-chapters describe these three learning objectives.

#### 3.2.1.5.1 LEARNING ABOUT THE PROCESS

Learning about the process means that humans gain a comprehensive understanding of the subject matter of decision-making. To be able to control the process, they need to develop profound expertise about both critical factors and how they interact with each other, as well as leverage points to interfere in a corrective or preventive manner. This allows the humans to monitor the system, develop situation awareness, detect problems, and find solutions (G. Klein & Wright, 2016). Learning about the process is a prerequisite to becoming an expert decision-maker.

#### 3.2.1.5.2 LEARNING ABOUT THE TOOL (AI)

Learning about AI-based tools means that humans gain knowledge of how AI functions as a tool. This is not focused on the AI's inner workings and algorithms, but rather on its capabilities and error boundaries, which humans need to understand to develop an accurate mental model (Bansal et al., 2019; Endsley, 2023b). This implies that humans are also aware of the AI's biases and potentially distorted views of the problem so that they know when they can rely on the AI's output and when they cannot. Learning about the AI tool is a prerequisite to obtaining appropriate trust.

#### 3.2.1.5.3 LEARNING ABOUT ONESELF

When working, humans show variability. For example, human decision-makers may tend to make riskier decisions towards the end of a shift. Learning about themselves means that humans gain a more comprehensive understanding of their behavior (Jelodari et al., 2023; Pronin, 2007) and can update their mental models of themselves. To support this, AI should provide transparency about humans' behavioral patterns and biases.

#### 3.2.1.5.4 EFFECTS OF AI APPROACHES ON HUMAN LEARNING

Table 9 provides an example illustrating how different AI approaches (Miller, 2023) are assessed regarding effective human learning according to the Experiential Learning Theory (Kolb, 1984). The Experiential Learning Theory is a cyclic process, which is characterized by a sequential learning progression, emphasizing the necessity of engaging with each stage in a systematic manner to ensure a thorough and effective learning experience. This sequence facilitates the conversion of experiences into actionable knowledge through a recurring cycle of experience, reflection, theory development, and experimentation (A. Y. Kolb & Kolb, 2009).

*Recommendations without explanations:* This approach does not engage with any of the four cyclic phases of experiential learning. While humans can have new experiences with AI, it is difficult for them to fully engage with AI if they do not understand AI or receive an explanation as to why AI has recommended something. Humans will have trouble understanding the recommendations without explanations. Without understanding, neither an accurate mental model of the task, the process, the AI, nor oneself can be created. Supporting the development of such mental models would be central to an effective human learning process.

*Recommendations with explanations:* This approach minimally helps humans to develop an accurate mental model of tasks and processes through the explanation of recommendations. However, because humans are not involved in developing the recommendations, it is difficult for them to understand and interpret the corresponding explanations. Consequently, humans can only make a few concrete experiences that reflect observations and support the construction of abstract conceptualizations of the task and the process. Furthermore, learning about the AI tool is limited, and learning about oneself is not supported at all.

*Recommendations with an interpretable model:* This approach somewhat helps humans develop an accurate mental model of the AI tool as well as of the task and process because the recommendations are more interpretable for humans. These interpretable models set the basis for humans to make concrete experiences with the task and process as well as with the AI tool. They can reflect on observations about the AI recommendations and thus consider what kind of recommendations an AI makes in different contexts. As a result, they can learn more about AI than with only explanations. They could learn even more about the tasks and the process when they get feedback about the utility and the actual success of the decision. Furthermore, learning about oneself is not supported at all.

*Cognitive forcing:* This approach enables people to have very concrete experiences by forcing them to deal with explanatory information (concrete experience). The cognitive forcing approach, given enough time, allows humans to reflect on their experience. Through this reflection, they can not only think about the AI, the task, and the process but also about their own behavior and thus also learn about themselves (reflective observation). This enables humans to build an accurate mental model of the AI, the task and the process, and even about themselves. Reflections allow humans to abstract their assumptions, develop new ideas, or adapt existing theories about the AI, the task, and the process, as well as about themselves (abstract conceptualization). However, this is only possible if the time, resources, and mental workload allow it. In this approach, humans are forced to actively experiment by applying abstract concepts in real-life scenarios (active experimentation).

*Evaluative AI:* This has a similar effect on human learning as cognitive forcing but goes further. Evaluative AI provides the human not only with explanations and recommendations on human-initiated decisions, but actively supports humans in exploring their own assumptions by providing evidence for and against these assumptions. This approach allows humans to go through all four stages of the experiential learning theory. In the process of hypothesizing and evaluating these hypotheses by the AI, humans gain concrete experience about the task and process as well as about the limitations of the AI tool. As the AI tool challenges the hypotheses proposed by humans, humans may even learn about themselves. This testing of hypotheses by the AI tool allows humans to reflect on their observations and view their experiences from many perspectives (reflective observation). Humans then conceptualize their reflections by developing new ideas or adapting existing theories (abstract conceptualization). Evaluative AI enables humans to use the new abstract conceptualization for their new hypothesis, which they test again with the AI. Thus, they actively experiment with abstract concepts in real-world scenarios and observe the results (active experimentation).

It should be emphasized that both cognitive forcing as well as evaluative AI involve *learning about oneself*, yet they only support this indirectly. In contrast, direct support could be achieved with an AI tool that observes human decision-making and provides direct feedback, which in turn stimulates human self-reflection.



Type of XAI (Miller, 2023)	Concrete Experience	Reflective Observation	Abstract Conceptualization	Active Experimentation
Recommendations without explanations	-/-	-/-	-/-	-/-
Recommendations with explanations	+/-	-/-	-/-	-/-
Recommendations with an interpretable model	+/-	+/-	-/-	-/-
Cognitive forcing	+/+	+/+	+/-	+/-
Evaluative AI	+/+	+/+	+/+	+/+

Note. Each type of AI (Miller, 2023) is evaluated to determine the extent to which the different learning phases of experiential learning (Kolb, 1984) are supported: -/- = no support; +/- = partial support; +/+ = fully supported.

**TABLE 9 – TYPE OF XAI RELATED TO THE CRITICAL PSYCHOLOGICAL STATES AND THEIR EXPRESSION**

**3.2.1.5.5 GENERAL CONCLUSIONS REGARDING AI4REALNET SCENARIOS**

In the context of the AI4REALNET project, particularly in relation to the second scenario of joint human-AI decision-making (including human-AI co-learning), it is becoming clear that this type of interaction between humans and AI is central to effective human learning with respect to all three learning objects - (1) learning about the process and the task, (2) learning about the AI and (3) learning about one's own behavior. This second scenario aligns with two types of AI support outlined by Miller (Miller, 2023): cognitive forcing and evaluative AI. These two types offer an approach through which people can gain a deep understanding of and an active engagement with AI tools.

In order to enhance human learning, the AI agent should be transparent and capable of communicating in an understandable way. This communication includes the AI application itself, the task and process, and the human behavior. Understandable explanations and transparency are crucial for humans to develop a deep understanding of all three learning objects and apply this knowledge effectively.

Moreover, AI should not only provide comprehensible explanations but rather support active human reflection or exploration in all four learning phases, according to Kolb (Kolb, 1984): Support in making concrete experiences, support in reflecting experiences, support in abstract conceptualization of reflections, and support in active experimentation with gained concepts. Ideally, this exploration is supported in relation to all three learning objectives, i.e., exploration of the task and the processes, exploration of the AI, and exploration of oneself.

AI4REALNET scenario one, i.e., AI-assistant to human (i.e., human in control), offers some opportunities to learn when it provides explanations and transparency. However, the learning support is limited as there is no interactivity, which prevents the important learning opportunity of exploration.

Finally, AI4REALNET scenario three, i.e., autonomous AI (human as a supervisor), does not support human learning at all.

### 3.2.1.6 HUMAN TRUST

Human trust in AI is one of the cornerstones of effective human-AI interactions (Jacovi et al., 2021; Lee & See, 2004). It represents a combination of beliefs, knowledge, emotions, and experiences that a person holds about an AI system (Cahour & Forzy, 2009). This complex construct influences how humans interact with, rely on, and integrate AI into their decision-making processes (Parasuraman & Manzey, 2010; Parasuraman & Riley, 1997). In contrast to trustworthiness, trust is not just a passive attribute, but a dynamic relationship between humans and AI that changes with every interaction (Hoffman, 2017).

The interaction between humans and AI is shaped by the specific *context of the tasks* at hand and the *capability of the AI* to perform these tasks, which crucially influences trust dynamics (Hoffman et al., 2018). Inappropriately high or low trust in AI can lead to significant mistakes: insufficient trust may result in neglecting useful AI recommendations and features, thereby missing out on potential benefits; too much trust can lead to over-reliance, potentially causing oversight of critical errors (Lee and See, 2004). The aim of the interaction between humans and AI should, therefore, be to enable humans to have appropriate trust in the AI. Appropriate trust mainly means having a realistic understanding of the AI tool's boundaries or scope of application (Miller, 2023). This means that, with increasing experience, a human should appropriately trust the AI tool for specific tasks or objectives in certain contexts or problem scenarios while also appropriately mistrusting the AI tool for other tasks or objectives in specific contexts or problem situations.

In the case of appropriate trust, automation transparency does not primarily mean that AI explains when and why humans should trust the AI tool's results. This is because explanations, for their part, presuppose blind trust in explanations. Rather, it means that humans can learn the limitations and capabilities of an AI tool through experience with it. If humans are not provided with the means to explore and test AI functionality, they might choose alternative methods or tools that they perceive to offer greater transparency and control (Koopman and Hoffman, 2003).

Experiencing the limits and possibilities fosters trust because people become familiar with the system through interaction. Therefore, the following challenges and requirements to gain appropriate trust are described.

#### 3.2.1.6.1 EFFECTS OF AI APPROACHES ON HUMAN TRUST

Table 10 assesses how different AI approaches (Miller, 2023) support the development of appropriate trust. Appropriate trust in AI emerges when humans learn through repeated interaction and experience to correctly assess the capabilities and boundaries of AI.

*Recommendations without explanations*, this approach fails to offer transparency and does not support the development of appropriate trust. It aligns poorly with developing appropriate trust because it provides no insight into the AI tool's capabilities and limitations, leaving humans unable to assess its reliability or relevance.

*Recommendations with explanations*: This approach provides explanations alongside recommendations, but this is not sufficient for gaining appropriate trust. Providing recommendations with an accessible level of complexity is very challenging. This challenge restricts humans' ability to gain experience and understand the boundaries of AI. Furthermore, the explanation of recommendations does not directly refer to AI's capabilities and limitations.

*Recommendations with an interpretable model:* This approach provides some insights into the inner workings of the AI and might partially reveal AI’s capabilities and limitations. However, although the model aims to be interpretable, without enabling exploration or offering full transparency regarding boundaries, for humans it remains difficult to thoroughly understand or evaluate the AI, which hinders the establishment of appropriate trust.

*Cognitive forcing:* Cognitive forcing provides a moderate level of exploration and transparency, thereby supporting the development of appropriate trust to some extent. By compelling humans to engage more deeply with the AI’s reasoning and decision-making process, this approach fosters a greater understanding and connection with the AI, laying a foundation for appropriate trust.

*Evaluative AI:* This approach stands out as the most effective in ensuring humans can gain appropriate trust. It excels in both exploration and transparency, directly involving humans in the decision-making process and providing clear insights into the AI tool’s functionality. By facilitating a deep and active engagement with the AI, evaluative AI empowers humans to critically assess and understand the AI’s capabilities and limitations, enabling the fostering of well-informed and appropriate trust.

Type of XAI (Miller, 2023)	Exploration	Transparency	Appropriate Trust
<b>Recommendations without explanations</b>	-/-	-/-	-/-
<b>Recommendations with explanations</b>	-/-	-/-	-/-
<b>Recommendations with an interpretable model</b>	-/-	+/-	+/-
<b>Cognitive forcing</b>	+/-	+/-	+/-
<b>Evaluative AI</b>	+/+	+/+	+/+

*Note.* Each type of AI (Miller, 2023) is assessed to determine the extent to which the various requirements for appropriate human trust in AI are supported. The evaluation is as follows: -/- = no support; +/- = partial support; +/+ = fully supported.

**TABLE 10 – TYPE OF XAI RELATED TO THE CRITICAL REQUIREMENTS AND THEIR EXPRESSION FOR ESTABLISHING APPROPRIATE TRUST IN AI**

**3.2.1.6.2 GENERAL CONCLUSIONS REGARDING AI4REALNET SCENARIOS**

In the AI4REALNET project, the development and implementation of different scenarios - AI-assistant to human (human in control), joint human-AI decision-making (including human-AI co-learning), and autonomous AI (human as a supervisor) - offer distinct pathways to establishing appropriate trust between humans and AI.

*AI-assistant to Human (Human in Control):* this scenario, characterized by recommendation-based decision support, can provide varying degrees of explanation and interpretable models. While this scenario provides a basic level of support, it inherently limits deep human involvement in the decision-making process. Although humans have an active role in deciding whether to accept or reject the decision proposed by the AI, they are completely passive regarding the generation of the decision. As described by many authors, this often leads to the so-called ironies of automation (Bainbridge, 1983) or ironies of AI (Endsley, 2023b): AI takes much more information into account in its decision-making

process than humans could. As a result, humans are usually overstrained when they have to decide whether to accept the AI's suggestion or not. This excessive demand can lead to the human either nodding off or rejecting the AI-generated suggestion. Both because they are unable to judge it. In contrast, if people are more actively involved in the actual decision-making process, they can better assess the capabilities and limits of the AI, which is a prerequisite for developing appropriate trust in the AI.

*Joint human-AI Decision-Making (Including Human-AI Co-learning)*, this scenario elevates the role of humans in the AI decision-making process, facilitated by approaches like cognitive forcing and evaluative AI. By integrating humans more actively, this approach significantly enhances the potential for evolving appropriate trust by:

1. Offering deeper insights into the AI's reasoning, fostering a better understanding of how AI conclusions are reached.
2. Granting humans a more active role, increasing the opportunity to explore and experiment with the AI tool to familiarize themselves with its boundaries.
3. Providing an avenue for direct feedback on the effectiveness of decisions, which is critical for validating the AI tool's accuracy (error boundaries) and utility.

In this collaborative model, trust is cultivated through a continuous loop of interaction and feedback, allowing humans to adjust their trust based on direct experience with the AI tool's performance.

*Autonomous AI (Human as a Supervisor)*, this scenario presents the most significant challenges for fostering appropriate trust due to no human involvement in the decision-making process. With humans relegated to supervisory roles, the opportunities for establishing a deep understanding and appropriate trust in AI are not given.

In essence, the key to cultivating appropriate trust lies in designing AI systems that are not only advanced in their technical capabilities but also in their ability to engage humans in a manner that promotes transparency, exploration, and feedback about performance and error boundaries. Such an approach ensures that trust in AI tools is informed by direct experience and a comprehensive understanding of AI's error boundaries, leading to more effective and nuanced human-AI collaborations.

### **3.2.1.7 ACCEPTANCE**

Cognitive engineering research has historically paid less attention to factors affecting the initial acceptance of new technology, thus factors possibly preceding trust, reliability, and others. Notice that the rejection of new technology can start at first exposure, perhaps even before an operator has used that technology. Notice a potential paradox in this: Operators might only develop trust after using a system, yet may be unwilling to trust a system they have not used. As such, initial acceptance of advanced decision-making automation can play a critical role in its successful deployment.

Sociology, psychology, and information systems communities, on the other hand, have studied factors underlying initial acceptance. Here, the compatibility between humans and technology is considered a key construct for overcoming the hurdle toward initial acceptance and technology adoption. "Compatibility," in this case, refers to the perceived fit of a technology within the context in which it is

used, driven by the user’s values, experiences, and needs. In general, the more compatible a technology is, the more likely it is to be accepted.

Human-machine compatibility can be found at various levels of cognitive work, ranging from basic handling qualities to decision-making styles and methods of task execution as illustrated in Figure 13. Research has shown that automated systems involving a high level of cognitive work are generally not well accepted amongst human operators. Empirical insights gained in ATC have shown that strategic conformance, the apparent strategy match between human and machine solutions, plays an increasingly important role in the acceptance of advanced decision aid (Westin et al., 2016). Similarity between human and machine solutions and/or actions is external, overt, and observable, and is the extent to which cause and effect can be observed.

In the ATC domain, the acceptance percentage of a personalized recommendation system (i.e., the system recommends human-like solutions to problems) was significantly better compared to a more general “one-size-fits-all” system (i.e., the system recommends more optimal solutions that are different from what human operators typically do). This result underlines the important role of strategic conformance in initial acceptance but also notes that, over time, the importance and practical benefits of strategic conformal automation can be questioned, considering daily and prolonged interaction with automated systems.

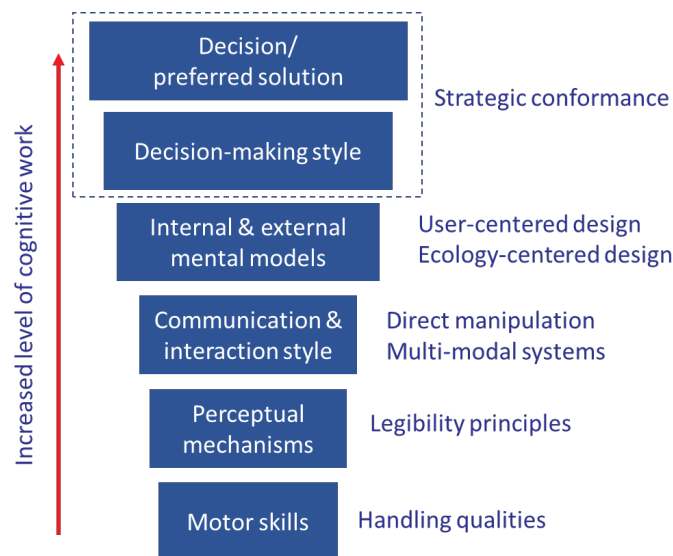


FIGURE 13 – LEVELS OF HUMAN-MACHINE COMPATIBILITY AND THEIR RESPECTIVE CONSTRUCTS FOUND IN COGNITIVE ENGINEERING RESEARCH ARE ORDERED BY INCREASED LEVELS OF COGNITIVE WORK; ADAPTED FROM (WESTIN ET AL., 2016)

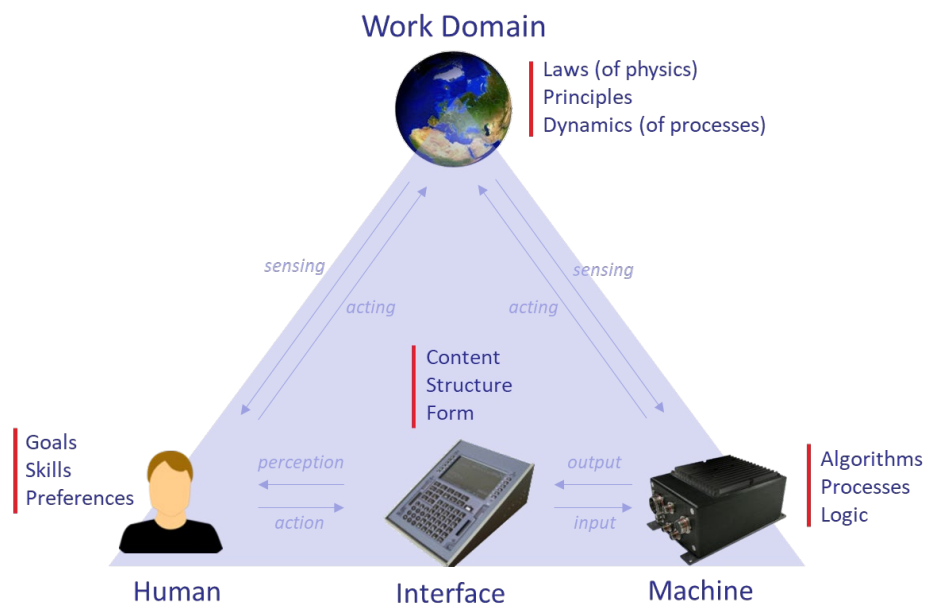
### 3.2.1.8 DESCRIBING AND DESIGNING HUMAN-AI INTERACTION

#### 3.2.1.8.1 TOWARDS A COMMON FRAMEWORK

For describing and designing human-AI interactions, lessons can be learned from human-automation interaction studies in cognitive engineering. These studies do not focus exclusively on AI, but on any form of technology with which human operators need to collaborate. In cognitive engineering, the gist of human-automation teamwork is centered around 1) team collaborations, with an emphasis on sharing and allocating control authority and autonomy between humans and automation, and 2) automation transparency, aimed at providing deeper system insights for fostering understanding,

trust, and acceptance. Currently, a generic design “cookbook” for human-automation interaction does not (yet) exist. Instead, AI4REALNET is exploring the integration of two promising and related *frameworks* that can be used for both analyzing and designing human-automation interaction: **Joint Control Framework** (JCF) (Vicente et al., 1995; Lundberg & Johansson, 2021) and **Ecological Interface Design** (EID) (Borst et al., 2015).

In its most succinct form, JCF focuses on team collaborations by describing the execution and planning of activities (e.g., sensing, deciding, and action implementation) when those are distributed over different agents. EID focuses more on achieving system transparency by visualizing the (physical and intentional) constraints on activities, which determine in large part the content, structure, and form of a human-machine interface. Integrating these two frameworks is possible due to their shared common ground: **Cognitive Systems Engineering** (CSE). CSE adopts a triadic approach to human-machine interaction where the design emphasis is first and foremost put on the work environment in which agents operate and activities take place – see Figure 14, where EID puts the emphasis on transparency by visualizing the constraints on activities, whereas JCF focuses on the execution and planning of activities (between elements). The work environment describes the boundaries for actions governed by physical laws, intentional principles, and processes. It essentially defines a safe envelope within which actions can take place, initially irrespective of who is executing the actions (e.g., human or automated agents). At later (design and analysis) stages, agent-specific constraints are included (e.g., capabilities and limitations of both human operators and machines).



**FIGURE 14 – TRIADIC APPROACH TO HUMAN-MACHINE INTERACTION.**

Given the shared CSE common ground, JCF’s emphasis on team collaborations, and EID’s focus on transparency, JCF and EID are complementary. The result of the first integration effort is shown in Figure 15. EID visually reveals the constraints, relations, and action opportunities at all functional abstraction levels, and JCF modulates human-automation coordination on activity level by putting (a sequence of) activities on a timeline describing on what abstraction level the system needs to be perceived, warranted by situational demands. In other words, EID prescribes what information should be portrayed and how, whereas JCF provides guidance on when to show information and how that

links to specific activities (e.g., perceiving system information, formulating a decision, performing an action, among others). In this figure, **D** represents a decision point, **AL** action leverage, and **PC** a perception point. **PC** represents the level of presenting information, e.g., in the figure, the status of the grid. A decision is, in this case, on the value level, e.g., the operator needs to decide on prioritizing serving customers or managing an overload. In this particular example, action leverage is on the plan level – this system has a means for describing a plan that can then be executed for the operator.

Consider the energy domain related to Figure 15: *1 Physical* is, e.g., breakers, lines, and their status. *2 Implementation* is, e.g., limits on performance such as voltage or current limits when operating a specific breaker. *3 Generic* is, for instance, a plan for solving an overload. *4 Value* represents trade-offs or limits; for instance, the need to serve/inform customers versus the need to resolve an overload can be defined by a specific voltage limit number. *5 Goals* represent what needs to be achieved, such as having a backup plan for possible forthcoming issues in the grid, serving customers, and avoiding overloads by looking ahead. *6 Framing* represents what is going on, on an overarching level, to manage a power grid, but specifically, what goes on in that management right now or in the future to be managed (e.g., an overload).

On the one hand, an AI that focuses on informing the operator would add perception leverage at higher levels and between levels. On the other hand, an AI that focuses on automating tasks would add action leverage at higher levels and between levels. An AI that is well-aligned either has Decisions, Action Leverage, and Perceptible Content on the same level or clear links between the levels so that relations between what is seen and what needs to be decided and done become clear.

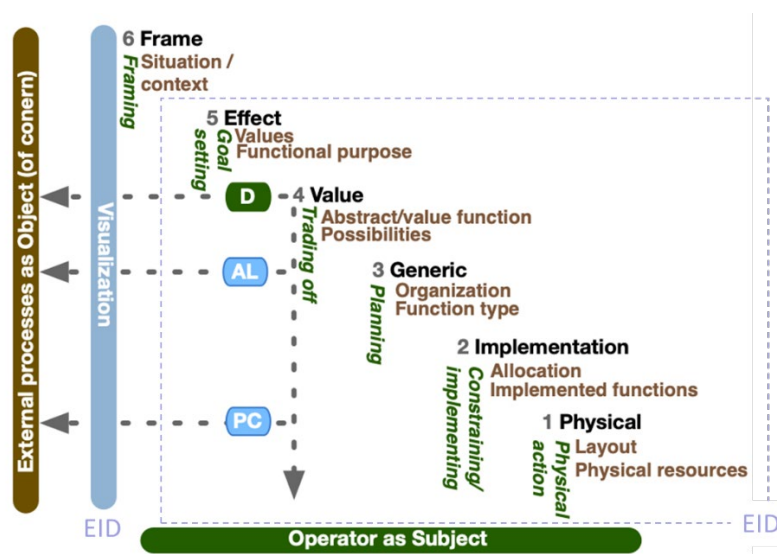


FIGURE 15 – MERGER OF JCF AND EID ON A FUNCTIONAL LEVEL.

3.2.1.8.2 AI4REALNET SCENARIOS

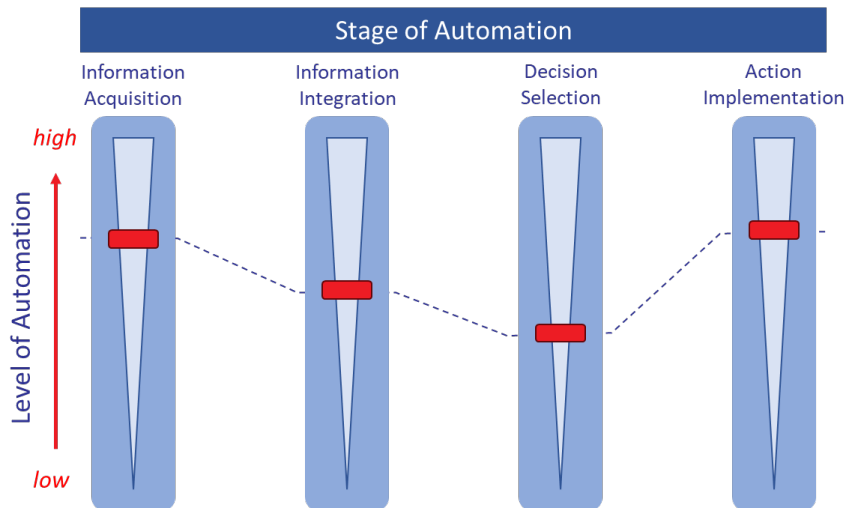
In AI4REALNET, three human-AI teamwork configurations are considered: **1)** AI-assisted human control (human in control), **2)** joint human-AI decision-making (including co-learning), and **3)** autonomous AI (human as supervisor). In cognitive engineering, these scenarios are embedded in the notion of “stages and levels of automation” (see Figure 16 below). At each stage, the levels of automation consider the division of roles and responsibilities between human and machine, and the delegation between the two, of both *autonomy* (i.e., how independently the system is permitted to initiate system changes)

and *authority* (i.e., the level of automation capability available to the system). The three AI4REALNET scenarios can be formulated in terms of the following stages of automation:

1. **AI-assisted human control:** AI features high levels of automation in *information acquisition*, *information integration*, and possibly *decision selection*. *Action implementation* is fully allocated to the human operator. A practical example is where AI directs humans' attention to important system information, integrates it in intuitive and human-friendly ways, and offers (a set of) directions where good decisions should be made.
2. **Joint human-AI decision-making:** AI and the human operator can both independently and autonomously observe information, make decisions, and undertake actions. In this configuration, bi-directional human-AI communication is required to ensure that both agents are aware of who is doing what, when, and how. A practical example is where the AI and human operator are working in parallel on completing a control task and, by observing each other's behavior, can learn from each other. For co-learning, it may be necessary to consider lower levels of automation at the *action implementation* stage, where the AI provides specific advisories that the human can accept, reject, or modify. This can be related to the project goals as follows, e.g.:
  - a. 1) the AI system could adapt continually to human preferences by analyzing
    - i. (1a) *explicit corrections made to its decisions, and 1b) implicit observations from the human's decision-making through his/her interaction with the user interface*. Examining Figure 2, the ability to carry out corrections by a human to make observations, and the ease of interpreting them by the AI, depend on the levels of interaction versus decision-making of the operator and AI.
    - ii. (1d) *typical preferences of the human operator in multi-objective problems*. For instance, regarding alternative plans (level 3), or priorities (level 4).
3. **Autonomous AI:** AI operates at the highest automation level at each stage, and the human operator needs to supervise the AI's behavior. Ideally, human operators do not need to step in, but in case of system faults, the human-AI system must fallback to lower automation levels and stages that allow human interventions.

It is important to note that choosing the right levels and stages of automation is warranted by operational contexts, situational demands, and capabilities (and limitations) of human and automated agents. As such, a "one-size-fits-all" distribution of functions and tasks does not exist and will need to be re-considered per application domain and/or operational scenarios.





**FIGURE 16 – STAGES AND LEVELS OF AUTOMATION MODELLED AFTER HUMAN INFORMATION PROCESSING STEPS (PARASURAMAN ET AL., 2000).**

In Figure 16, four activities can be seen that are related to the interface information described in Figure 15. The first step, IA relates to using perceptible content. In JCF, this corresponds to a *perception point*, P, in the operator process of information pick-up and use. The second stage, *information integration*, has to do with preparing the information so that it matches what is needed for decision selection. For instance, it can be presented on a different abstraction level, matching what the operator needs. The third step, *decision selection*, matches the *action leverage* in Figure 15 and, when carried out, represents a *decision point* (D) in JCF. The third part, *action implementation*, corresponds to an *action leverage* in Figure 15 and to an *action point* (A) in JCF.

In Figure 16, we also see an arrow, going from high to low, denoting a concept of “level of automation” (LOA) within a stage of automation. The LOA denotes how independently an operator or an automated system works with that information. The extremes (high/low) usually denote fully manual or fully automated. In Figure 16, the important point is that the LOA can differ regarding these four stages of completing a cycle of gaining information and acting on it. Various academic proposals on LOA have been presented, and moreover, applied fields have their own LOAs. The three domains in AI4REALNET can choose to use an application-field-specific LOA or a generic one from academia. Using a generic LOA facilitates cross-domain comparisons.

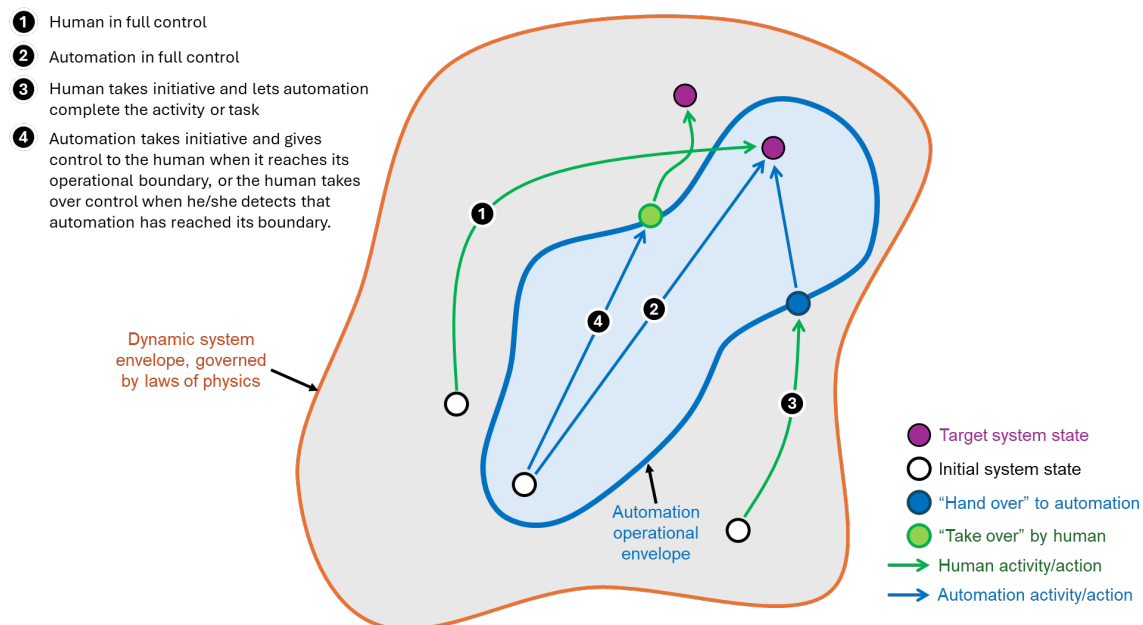
### 3.2.1.8.3 GENERIC OPERATIONAL EXAMPLE

To illustrate the combination of JCF and EID, consider Figure 17 showing an abstract state-action space of a generic planning problem where the goal is to bring a system from an initial state toward a desired target state. Generally, safe and unsafe actions within a planning problem are bounded by causal laws and intentional principles that are independent of any particular agent that can execute actions. Human agents typically “satisfice” by performing safe actions to reach target states, but human actions are seldom optimal. In contrast, automated agents aim to “optimize” by taking the shortest possible route toward the target state. However, automation may have a limited operational envelope for executing the shortest route, requiring human operators to oversee the automation’s functioning, anticipate when it reaches its boundary and timely take over control when that boundary is crossed. On the other hand, humans can also collaborate with automation, for example, by taking the initiative and ‘hand over’ a task to automation when the system state is within the automation’s operational

envelope. Note that automation can be classical, rule-based automation, or AI-based learning automation. For rule-based automation, the operational envelope is often fixed, whereas the envelope for learning-based automation can adapt by learning from (past and new) experiences (e.g., human actions, historical data, etc.).

Traditionally, in EID, the emphasis lies on discovering and portraying operational envelopes governed and bounded by laws of physics. Within an ATM perspective, this relates to, for instance, the turn and climb performances of an aircraft and how certain physical objects (e.g., other aircraft) may pose restrictions (or obstacles) within state-action spaces. Central in EID is Rasmussen’s Abstraction Hierarchy (AH) – an actor-independent and activity-independent functional map that describes the overall system at different abstraction levels, ranging from the system’s functional purpose to its physical form. The AH typically specifies the content and structure of an interface, and the goal is to portray the AH information to transform a cognitive task into a perceptual task.

In general, portraying system envelopes does not dictate a specific course of action but empowers the human operator to take any action as long as it does not violate the constraints. When human operators need to collaborate with an automated agent that operates in the same work environment (and thus is bounded by the same natural laws), insights into the automation’s operational envelope become important, as well as coordinating activities between agents (as illustrated in Figure 17). In this regard, JCF complements EID by describing and analyzing sequences of activities/actions on a timeline, at what abstraction level agents (need to) perceive system information to coordinate activities/actions, and what control authority each agent (needs to) has. Such information is crucial in analyzing the system's stability in terms of patterns in human-automation interaction.



**FIGURE 17 – ABSTRACT STATE-ACTION SPACE DESCRIBING A GENERIC PLANNING PROBLEM WHERE HUMANS AND AUTOMATION CAN COLLABORATE (IN SERIAL OR PARALLEL) TO BRING THE SYSTEM FROM AN INITIAL STATE TOWARD A SAFE TARGET STATE (VAN PAASSEN ET AL., 2018).**

To understand how to use the JCF tools in analyzing human-automation interaction patterns on the activity level, consider again Figure 16, which presents an abstract set of situations encompassing teamwork where humans and automation can work either in serial or parallel. When humans and

automation work together in parallel, scenario ① in Figure 17 represents the situation where humans are in full control, whereas scenario ② represents the situation where the human must monitor the automation that can take optimized actions. When working together in serial, the automation may provide optimized recommendations that a human needs to inspect and evaluate, consequently accepting, revising, or rejecting the automated advice. After rejecting the advice, the human becomes responsible for formulating and executing an alternative action. Other teamwork organizations may involve coordinating “handovers” ③ and “take-overs” ④ between agents. For example, the human could bring the system to a desired target state by formulating a plan but hand over the execution of that plan to automation (③). Vice versa, the automation could also formulate a (partial) optimized plan but hand over the execution of that plan to the human (④). Note that scenario ④ could also represent a situation where automation is not able to bring the system to the desired state (e.g., due to unpredictable weather conditions that fall outside the automation’s operational envelope), requiring the human operator to take over control.

While EID helps in specifying what information needs to be shown and findings ways to show that information, each scenario described above has an impact on the allocation of control authority and autonomy between agents, how activities are or should be coordinated, and when what type of information is or needs to be accessed. The JCF provides two tools for describing this: The Level of Autonomy in Cognitive Control (LACC) – Level of Automation (LOA) matrix and the JCF-Score. The main advantage of this is that it facilitates cross-domain comparisons between the AI4REALNET cases. This also provides a backdrop for discussing the generalizability of solutions across cases.

The JCF offers a way to systemically describe stages of automation and link them to information requirements found at various functional abstraction levels using the LACC-LOA matrix, see Figure 18 (which is just an example, many variations and options are possible in AI4REALNET). Note that in the conditional automation case of this figure, if the human plans and optimizes, it becomes identical to the scenario ① case – but there is a distinction. In scenarios ② and ③, in this figure, the monitoring task is an added human activity that is not present in scenario ①. Thus, scenario ① is within one operational envelope, scenario ② within another, and scenarios ③ and ④, cross the envelope borders.

<i>Framing</i>	<b>FRAME</b>	FRA	1A	FRA	1B	FRA	1B	FRA	2A
	<b>Automation situational variety</b>	<i>understanding the automation limits</i>		<i>planning and executing</i>		<i>hand-over or joint planning</i>		<i>planning and executing</i>	
<i>Situation/context</i>									
<i>Goal setting</i>	<b>EFFECT</b>	EFF	1A	EFF	1B	EFF	1B	EFF	2A
	<b>Automation goal variety (targeting)</b>	<i>understand side-effects to communicate to customers</i>		<i>goals are pre-set</i>		<i>goals are pre-set</i>		<i>goals are pre-set</i>	
<i>Functional purpose</i>									
<i>Trading off</i>	<b>VALUE</b>	VAL	1A	VAL	1B	VAL	1B	VAL	2A
	<b>Automation value variety</b>	<i>understand KPIs and make trade-offs eg between serving customers efficiently and having safety margins</i>		<i>human optimizes</i>		<i>human optimizes OR adjusts optimization from automation</i>		<i>automation optimizes</i>	
<i>Abstract/value function</i>									
<i>Planning/monitoring</i>	<b>GENERIC</b>	GEN	1A	GEN	1B	GEN	1B	GEN	2A
	<b>Automation functional variety</b>	<i>understand affected plans</i>		<i>human plans</i>		<i>human plans OR adjusts plans from automation</i>		<i>automation plans</i>	
<i>Function type</i>									
	<b>IMPLEMENTATION</b>	IMP	1A	IMP	1B	IMP	1B	IMP	2A
<i>Constraining / implementing</i>	<b>Automation implementation &amp; execution variety</b>	<i>understand when to act</i>		<i>automation gives guidance</i>		<i>automation gives guidance</i>		<i>automation gives guidance</i>	
<i>Implemented functions</i>									
<i>Physical action</i>	<b>PHYSICAL</b>	PHY	1A	PHY	1B	PHY	1B	PHY	2A
	<b>Automation action/ object variety</b>	<i>observe cues for automation limits</i>		<i>automation implements guidance or steers</i>		<i>automation implements guidance or steers</i>		<i>automation implements guidance or steers</i>	
<i>Physical resources</i>									
	<b>Levels of Autonomy In Cognitive Control</b>	Human in full control		Divided work: Human plans, automation executes		Conditional automation		Automation in full control	
		<u>scenario 3, 4 monitoring task</u>		<u>scenario 1</u>		<u>scenario 3,4 planning and execution task</u>		<u>scenario 2</u>	

FIGURE 18 – LACC-LOA MATRIX FOR THE EXAMPLE IN FIGURE 17.

In Figure 19, the temporal execution of scenario ③ can be seen in the JCF score notation. The Score has two main processes: the bottom is human-in-control, and the top is automation-in-control. It has a tentative temporal distribution (to be empirically set for any case that uses this pattern) on the horizontal axis. On the vertical axis, the numbers 1 – 6 represent the level of abstraction, originating from EID, at which information needs to be accessed (e.g., one typically represents the physical form, related to the topology of objects and their status). In the case of an aviation example, the activity pattern first starts with an observation of aircraft status and destination, then recognizing that a plan is needed. Optimization needs are decided, and a plan is made and entered into the system by the human. Note that this same problem may occur in the energy domain, where an overload status would instead be observed, and grid optimization needs would be decided, and a human could enter a plan for resolution. When the plan has been entered, the automation observes this and gives guidance to the execution layer below (e.g., regarding the timing of actions), and then the automation implements this guidance. In the aviation case, it gives clearances; in the grid case, it operates switches; in the rail case, it operates railway lights and rail switches.

It strips away the form of interactions but describes the content and LACC level of interactions over time. The exact timings are an empirical question; this score describes a tentative case and roughly denotes the order of interactions, as well as the important crossing between human and automated work. The timings can also be *designed* for a particular domain – timings that must later be empirically validated.

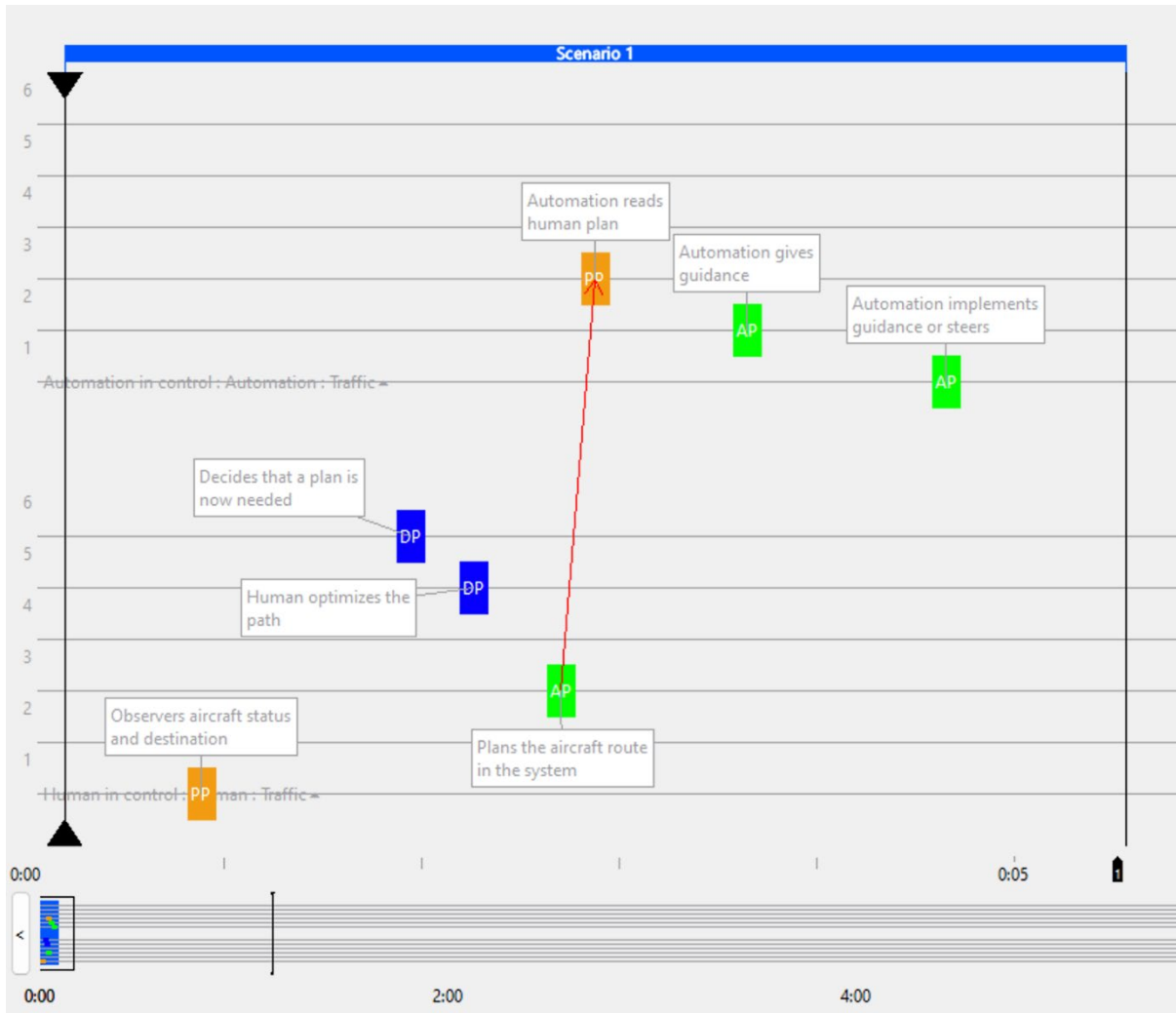


FIGURE 19 – JCF SCORE FOR SCENARIO 3

### 3.2.2 AI AGENT AND DECISION MAKING

AI-based decision-making is increasingly transforming the landscape of human-AI collaboration, offering unprecedented capabilities in processing complex data, identifying patterns, and generating insights that surpass human cognitive limits. In the context of human-AI decision-making, AI systems can augment human judgment by providing data-driven recommendations, enhancing efficiency, and reducing bias in critical decisions. However, this synergy also brings challenges, including the need for transparency, trust, and ethical considerations to ensure that AI supports, rather than undermines, human autonomy and values. Balancing the strengths of AI with human intuition and expertise is essential to harness the full potential of AI-based decision-making in a responsible and effective manner. The objective of this chapter is firstly to elaborate on the different characteristics that AI-based models should possess for their integration into the AI4REALNET framework. This also implies that these characteristics should allow efficient interactions between AI and Human decision-makers in various situations and modes of interaction introduced earlier. Additionally, some methodological and algorithmic aspects of AI-based models are introduced.

### 3.2.2.1 CHARACTERISTICS OF AI-BASED DECISION MODELS

#### 3.2.2.1.1 ROBUSTNESS, RELIABILITY AND RESILIENCE

Reliability in the context of AI and ML systems simply refers to the basic ability of the model/algorithm to perform as intended over a specified time frame under specific conditions, ISO/IEC TS 5723:2022. On the other hand, robustness goes beyond this standard situation to consider the ability of a system to maintain its level of performance under a variety of circumstances, ISO/IEC TS 5723:2022. When considering AI models, robustness can be categorized into two types: *algorithmic robustness*, pertaining to the sensitivity of the learning algorithm to perturbations in its training dataset, and *model robustness*, which describes how a trained model reacts to perturbations in the input data. In a final step, resilience can be considered the robustness of the AI model with regard to security threats. In other words, resilience constitutes the ability of an AI system to prevent, respond to, and recover from adversarial attacks.

In the academic literature, the verification of AI/ML-based systems has predominantly considered computer vision problems with artificial neural networks of different architectures. According to a recent literature review by (Ilahi et al., 2021), the number of publications and methodologies that study the impact of adversarial attacks in deep learning algorithms that do not use images as inputs is low. For RL, the authors defined four categories of adversarial attacks targeting 1) state space, 2) reward function, 3) action space, and 4) model space. In supervised and unsupervised learning, type (1) is ‘input space’ instead of ‘state space’, type (3) is ‘model output’ instead of ‘action space’, and (4) applies to any learning paradigm.

For critical infrastructures and the six UCs of the AI4REALNET project, the risk qualitative assessment of Table 11, based on the dimensions of ETSI GR SAI<sup>15</sup> 001, can be applied to identify where the focus should be placed in terms of adversarial or natural perturbations.

		Attack targeting/Failure on			
		Model space	Reward function	Action space	State space
Magnitude	This can lead to grid outage events, congestions (delays) in the railway network and ATC spaces, non-optimal economic control solutions, or high carbon emissions (e.g., excessive curtailment of renewable generation), with monetary and reputation loss and a negative impact on the economy and comfort levels.				
Duration	Reward functions and models are generally stored and operated in highly cyber-secure Information Technology (IT) systems. In the event of an attack, the previously trained model could be quickly restored.		Associated with the Operational Technology (OT) systems, which follow high cybersecurity and reliability standards. Moreover, a lack of knowledge about network topology and parameters makes attack duration difficult.		Data-driven models are often vulnerable to small imperceptible perturbations to the input data (Goodfellow et al., 2014). Events such as missing or erroneous data can be common in real-world networks.

<sup>15</sup> Securing Artificial Intelligence (SAI). AI Threat Ontology. ETSI GR SAI 001 V1.1.1 (2022-01). [Online] [https://www.etsi.org/deliver/etsi\\_gr/SAI/001\\_099/001/01.01.01\\_60/gr\\_SAI001v010101p.pdf](https://www.etsi.org/deliver/etsi_gr/SAI/001_099/001/01.01.01_60/gr_SAI001v010101p.pdf)

	Attack targeting/Failure on			
	Model space	Reward function	Action space	State space
Scope	National and regional impact, such as overloaded power lines, cascading failures, load shedding, and train and flight delays.			
Severity	Human operators must decide without AI assistance or an autonomous AI system (transfer of control to human). Decrease in trust over AI.			
Response	Previously trained models could be quickly restored. Model training is done in a secure and controlled digital environment, and model retraining is possible.		Model replacement or re-training does not solve the problem. During operation, it is primarily a cybersecurity issue. Model training is done in a secure and controlled digital environment (or twin).	Model re-training is not possible during operation. Model replacement does not solve the problem.

**TABLE 11 – AN EXAMPLE OF RISK QUALITATIVE ASSESSMENT OF THE UC BASED ON THE DIMENSIONS OF ETSI GR SAI**

In the critical network infrastructure context, the focus is on perturbations in the state/input space under natural or adversarial changes in the observations. Note that if the digital environments, or twins, accurately emulate operational scenarios of real-world networks and events, the focus would be on normally trained AI-based systems and controllers, referred to as ‘test-time’ by (Behzadan and Munir, 2017). However, changes in the concept, data, external systems, or software pipeline can result in out-of-domain data and/or data drift that may significantly decrease the AI system’s performance.

Evaluating the robustness, reliability, and resilience of AI systems during training and testing time is paramount in critical infrastructures. Consequently, a formal definition of these concepts is presented in the following, derived from harmonizing current AI taxonomy harmonization, standards, and academic literature.

**Robustness**

The robustness of an AI system encompasses both its technical and social perspectives (EU-U.S. Terminology and Taxonomy for Artificial Intelligence<sup>16</sup>).

*Technical robustness* is a system’s ability to maintain its performance level under natural or adversarial perturbations. It can be local (specified with respect to a sample input) or global (guarantees that hold deterministically over all possible inputs), according to ISO/IEC 24029-2. Note that considering the complexity of the systems at hand in AI4REALNET, local robustness is easier to specify and verify. This ability can be evaluated using two methods. The first utilizes the sensitivity property (ISO/IEC 24029-2) that measures the extent to which the output of the AI system or the reward/loss function varies when its inputs are changed, where metrics such as output/reward variance can be used. In the second method, an adversarial agent applies perturbations to the AI system, replicating natural and intentional scenarios (which can be imperceptible perturbations), where the difference between the total rewards/loss obtained with the unperturbed and perturbed systems is a potential metric for robustness. The adversarial agent can also be used to quantify the sensitivity property. The range of

<sup>16</sup> EU-U.S. Terminology and Taxonomy for Artificial Intelligence. First Edition. [Online] <https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence>

change in the output (action space) can serve as a metric for the technical robustness of the system, for example by assessing whether a particular decision holds for input variation (noise, missing data) in the same context. Furthermore, during training-time, the magnitude of the reward/loss function deterioration can be used to measure robustness (Behzadan and Munir, 2017). Stress tests with these metrics are necessary for different perturbation probabilities, a maximum number of perturbations or a perturbation budget. This should be properly crafted in the adversarial agent reward function. Finally, the detection of out-of-domain data/data drift differs from technical robustness assessment to external perturbations or events. It is inherent to the model and learning mechanisms. For instance, the use of online learning changes the AI system’s behavior, which can also change its robustness (in a positive or negative direction). This also means that test-time robustness monitoring is needed on a regular basis for AI systems that use online learning.

*Social robustness* should ensure that the AI system duly considers the context and environment in which it operates. The ALTAI framework can guide end-users in this assessment and lead to new functional and non-functional requirements. Moreover, in the AI4REALNET concept, digital environments play an important role by simulating the impact of a perturbed AI system with KPIs of social relevance, e.g., carbon emissions reduction of the power grid (see the KPI list in section 2.3).

### **Reliability**

According to the EU-U.S. Terminology and Taxonomy for AI, “an AI system is said to be reliable if it behaves as expected, even for novel inputs on which it has not been trained or tested earlier”. This definition is strongly related to out-of-domain data. In other words, this means that the AI system should perform similarly on any test sets/periods if they are from the same distribution. This is closely related to the concept of generalization, as discussed in the section 3.2.2.1.4.

In contrast to robustness, which considers the influence on the performance of AI system operating context (e.g., natural or intentional perturbations, faults in the subsystems such as forecasting functions), reliability focuses on consistent performance aligned with the underlying data distribution in standard operating environments (Zissis, 2019). Estimation of epistemic uncertainty (discussed in section 3.2.2.1.6) provide valuable information, correlating model performance with the level of uncertainty. Models with better performance in areas with high epistemic uncertainty can be considered more reliable.

### **Resilience**

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<sup>17</sup> (EU-U.S. Terminology and Taxonomy for AI). Here, it is important to highlight the notion of recovery in resilience.

Its quantification is related to the magnitude and/or duration of reward/loss function performance degradation compared to an unperturbed system for the same context. Figure 20 depicts a conceptual definition of the resilience quantification for a reward function. In this scheme, resilience can be quantified by a) the grey area between the reward curves of the unperturbed and perturbed AI system,

---

<sup>17</sup> According to NIST AI 100-1, “security and resilience are related but distinct characteristics. While resilience is the ability to return to normal function after an unexpected adverse event, security includes resilience but also encompasses protocols to avoid, protect against, respond to, or recover from attacks. Resilience relates to robustness and goes beyond the provenance of the data to encompass unexpected or adversarial use (or abuse or misuse) of the model or data”.



b) minimum reward in the degradation state and maximum reward in the restorative state, and c) duration of the degradation and restorative stages. These metrics should be computed for different probability levels of the perturbations or by defining a maximum number of perturbations or a perturbation budget.

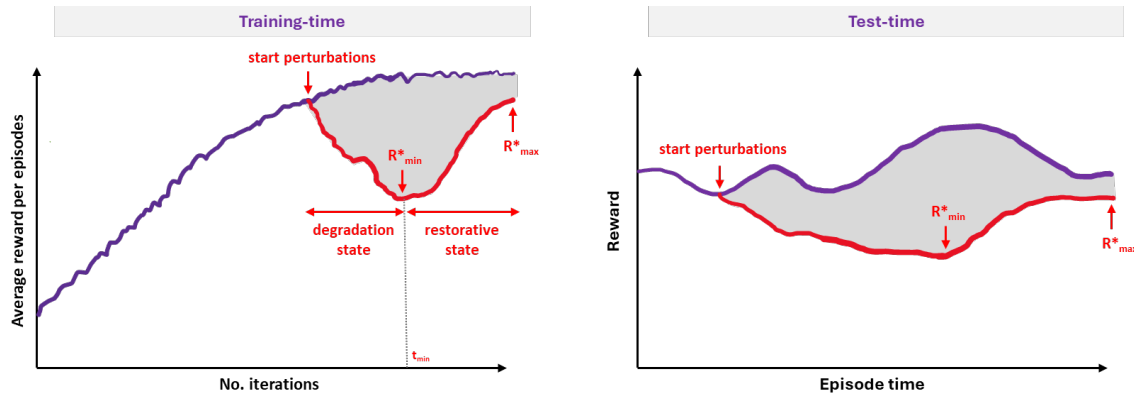


FIGURE 20 – CONCEPT OF RESILIENCE QUANTIFICATION IN TRAINING-TIME AND TEST-TIME PHASE

### 3.2.2.1.2 INTERPRETABILITY AND EXPLAINABILITY

Explainability measures the capability of a human user to understand how models make predictions or decisions, where the model’s transparency is a way to support explainability (Miller, 2023). According to (Molnar, 2020), explainability is contextualized by a specific input, and it often requires additional information, which is not generally generated by the decision model. XAI techniques refer to the set of methods that aim to generate local explanations for black-box model’s, e.g., Artificial Neural Networks (ANNs) predictions.

**Explainability** is crucial in decision-making as it fosters trust and acceptance of AI models by human users (Gunning & Aha, 2019). In RL, explainability addresses the challenge of understanding the long-term impact of a certain decision, a task that humans find difficult to grasp as they tend to perceive the immediate reward to outweigh future rewards, as detailed by (Loewenstein & Elster, 1992).

As deep ANNs have advanced, the field of RL has undergone a significant transformation. The shift from simpler learning representations to ANNs has empowered AI agents to tackle tasks that were otherwise not addressable. However, this transition led to these convoluted models being treated as black boxes. Consequently, there is an urge to develop methods that can bridge the gap between human understanding and AI decision-making, ensuring that decisions made by RL agents are both transparent and trustworthy for users, as highlighted in (Li, 2017). The lack of explainability and transparency often hampers the deployment of RL in real-world applications.

For a given decision, there are often multiple plausible explanations. The authors of (Heuillet et al., 2021) suggest tailoring explanations to the targeted audience and its goal. Therefore, the AI agent developed within the project should support multiple means of explanation to accommodate the different actors who may be interested in an explanation. These actors range from the operators, who interact with the system, to regulatory agents. The latter comprises organizations and agencies responsible for ensuring that the system is compliant with the standard defined for the addressed domain.

Given the targeted domains of critical infrastructure, explainability plays a pivotal role in guaranteeing that an AI agent does not pose new threats. As an example, when AI is deployed as an assistant to humans, an operator may ignore the received suggestion that is not fully understood. As a result, a well-explained sub-optimal decision might be preferred over a superior option that lacks a clear explanation.

Among the comprehensive set of KPIs, requirements, and metrics employed across the identified UCs (see section 2.3), those related to human trust and acceptance are the most relevant to the topic of explainability. In addition, it is crucial that the computation time required to formulate an explanation, along with the processing time a human requires to decode and understand it, do not delay the decision-making. Finally, explanations should be assessed for *fidelity*, which measures the accuracy of an explanation in representing the underlying decision-making process, contextualized by the current input.

To ensure a thorough evaluation, (Vouros, 2022) proposes a set of additional human-related metrics evaluating the interaction between a human and the explanation. These metrics, are reflected in the broader list of KPIs for each UC in Section 2.3, focus on the effectiveness of the explanation from a user’s perspective. The metrics relevant to explainability include:

- **Comprehensibility:** assessing the capacity of a human to understand an explanation.
- **Preferability:** estimating the relevance of an explanation given to the user.
- **Cognitive load:** estimating the cognitive effort required by a human to appreciate and comprehend the provided explanation.
- **Actionability:** assessing the utility of an explanation by capturing how well an explanation enables end-users to make informed decisions.

Considering these metrics from an early stage of development allows for the design of a human-friendly AI, where humans can be in control as it enables users to understand the system’s decision.

#### 3.2.2.1.3 TRACEABILITY AND AUDITABILITY

**Auditability** consists of a thorough analysis of data, algorithms, and design processes to ensure alignment with the desired objectives, standards, and legal and technical requirements, such as those outlined by the European Commission (EC, 2024). This concept is pivotal in building human trust in AI systems. One might argue that, when deploying AI/ML into real-world systems, auditability is as important as model performance.

To ensure auditability, a continuous process that evolves throughout the entire lifecycle of the AI system is required. Significant updates to data, architecture, or system design (Markov Decision Process – MDP – for an RL application) require a careful re-evaluation of auditability. Logging and tracing from an early stage of the AI system design and development are essential to ensure auditability. These mechanisms ensure the correct recording of any meaningful and relevant insight that may be essential to ensure trust in the AI system from human users.

In the context of software development, **traceability** refers to the process of establishing a clear and direct connection between the stakeholders’ requirements and the product developed<sup>18</sup>. When applied to AI/ML systems, this connection covers the design elements, code implementation, test

---

<sup>18</sup> Traceability. <https://en.wikipedia.org/wiki/Traceability>. Accessed: 3rd April 2024.

scenarios, and data used to train the system. To ensure continuous traceability of RL in human-centered AI, the human-machine interaction, along with the corresponding context, must be logged to trace the human influence in future decisions.

In the AI4REALNET project, across the six UCs, auditability and traceability have a three-fold goal. Firstly, in such critical domains, traceability and auditability are essential to guarantee that the human is in control and the network can be safely operated. In the context of RL, this means implementing automated controls to detect changes in any MDP element, such as observation or action space. These controls guarantee that the task addressed remains consistent, guaranteeing reproducibility under the same circumstances and preventing the policy from being misled by external influences. Secondly, auditability and traceability are crucial in maintaining control and safety, ensuring regulatory compliance, and allowing for effective monitoring and inspection of both AI recommendations and human inputs. Thirdly, auditability and traceability should work in two ways: to trace AI recommendations and human inputs, and to allow for quantification of the human influence over the AI-powered decision system. Finally, auditability and traceability are helpful for identifying and investigating performance degradation. By pinpointing the exact causes of issues, these processes significantly enhance the system's maintainability.

#### 3.2.2.1.4 GENERALIZATION

Generalization, in the context of AI/ML, is the ability of a trained model to perform well on previously unseen data that are derived from the same distribution as the data explored during the training phase. For RL, (Nichol et al., 2018) consider an agent to generalize well when it can adapt to previously unencountered situations drawn from the same MDP explored during the training phase, e.g., different levels of a game. Nonetheless, (Cobbe et al., 2019) claim that generalization in RL is still an open challenge as state-of-the-art algorithms are generally trained and evaluated within a limited set of tasks.

The generalization problem in RL may be related to the problem of overfitting, and, as argued by (Irpan, 2018), this may be due to the policy being optimized based on the reward signal. Generally, in RL, the agent's behavior is shaped by a reward signal, which is often formulated ad-hoc for specific scenarios. Consequently, if the reward formulation is narrowly tailored to one case, the resulting policy may underperform in other scenarios.

(Zhang et al., 2018) suggests that generalization in RL can be improved by letting the agent visit multiple diverse instances during the training phase. Nevertheless, a small perturbation of the environment may still hinder the agent's capabilities to accomplish the task.

The problem of generalization in RL should be addressed on three different levels:

- **Domain diversity:** An agent must visit a variety of environment configurations to promote the exploration of the state-action space and to lower the chances that an agent will stick to a restricted sequence of actions leading to the fulfillment of the goal by exploiting the determinism of the environment.
- **Exploration-exploitation trade-off:** An agent must balance exploration and exploitation to prevent overfitting to certain cases or local optimum.
- **Experience diversity:** The learning model underneath the agent must be optimized on various experiences to prevent biases introduced by the optimization on a restricted set of data (Olteanu et al., 2019).

To prevent an RL model from overfitting to a narrow area of the state-action space, prior work has focused on improving the exploration phase of RL to improve the model's robustness. *Curiosity-driven* methods consist of augmenting the reward signal with a value that encourages exploration of yet unknown sub-areas of the observation space (Pathak et al., 2017; Li et al., 2020).

In the context of the AI4REALNET project, we have established a set of requirements to ensure that the AI model is robust and capable of being generalized across diverse scenarios. Where possible, the trained model should maintain its performance in unseen and out-of-distribution scenarios, such as areas of observation and action space not visited during training. If this is not feasible, the AI agent must alert the operator about its uncertainty as outlined in section 3.2.2.2.2. The generalization capability of a model could be measured by observing the change in the reward/loss of the model when visiting novel data.

#### 3.2.2.1.5 SCALABILITY

In AI, scalability is the ability of a model to adapt to different workloads, similar to the algorithm's scalability as outlined by (Paliouras, 1993; Ulanov et al., 2017), where they assess the scalability of an AI/ML distributed model by measuring the empirical speedup obtained from a system while increasing the computational resources.

In each of the six UCs of the AI4REALNET project, an RL-powered AI agent will be used to either provide recommendations to a human operator or in a fully autonomous manner. As a result, the scalability challenge is two-fold. On one hand, from an engineering perspective, an AI decision-making model should scale based on the hardware availability. On the other hand, from a theoretical RL perspective, the system's effectiveness and performance should not be compromised by the level of complexity given by the combinatorial nature of Multi-Agent RL (MARL) with an arbitrary number of agents (Hernandez-leal et al., 2019). In MARL, the decentralized decision-making process preserves the integrity of the MDP-based learning strategies even as the system scales. As an example, in the context of the Flatland digital environment defined by (Mohanty et al., 2020), the complexity grows exponentially with the deployment of additional trains on the rails and with the railway network expansions. Consequently, the system performance, measured through time elapsed for decision-making and the model's accuracy, may be affected by the additional agents.

The critical infrastructures addressed within the project generally require immediate intervention to keep the network in its normal operational status by addressing unforeseen critical situations that may arise. A decision from the human-AI team must be made in near real-time to prevent an issue's escalation, regardless of the scale of the problem or the complexity of the network. Consequently, it is crucial to consider the scalability constraint from an early stage of the design phase. The training and inference methods, along with the algorithms, must be designed to accommodate large and realistic scenarios. This could be achieved through MARL by factorizing the learning process across multiple agents such that each agent can learn and make decisions simultaneously within a shared environment.

#### 3.2.2.1.6 UNCERTAINTY QUANTIFICATION

Uncertainty quantification (UQ) is a critical component when integrating AI into decision-making processes for critical infrastructures. This approach involves systematically characterizing and managing the uncertainties inherent in both the AI models and the real-world data they process.

In the context of human-AI interaction, UQ ensures that decisions made with the aid of AI are reliable and robust. There are several aspects to UQ:

1. **Model uncertainty (epistemic):** AI models are not infallible. They are built on algorithms and data that might not always perfectly capture the complexities of the real world. UQ helps identify the confidence level of AI predictions, highlighting areas where the model's output is less certain. In the context of the human-in-the-loop pipeline, which will be used in the AI4REALNET framework for decision-making, the estimation of epistemic uncertainty may allow the AI agent to establish its level of confidence within an observed state.
2. **Data uncertainty (aleatoric):** The data fed into AI systems often comes with its own uncertainties due to noise, incompleteness, or inaccuracies. In the context of the human-in-the-loop pipeline, an environment should support the estimation for aleatoric uncertainty derived from an external source, such as weather conditions. UQ methods, such as probabilistic modeling, can quantify these uncertainties, providing a clearer picture of the data's reliability.
3. **Decision-making under uncertainty:** For critical infrastructures, decisions must be made with an understanding of the potential risks and outcomes. UQ supports this by offering a probabilistic framework that can be used to evaluate different scenarios, helping human operators to make informed decisions even in the face of uncertainty.
4. **Human-AI collaboration:** UQ fosters better collaboration between human decision-makers and AI systems. Providing transparency about uncertainties allows humans to apply their judgment effectively where the AI's predictions might be uncertain or ambiguous. As an example, for the power grid use cases of the AI4REALNET project, the AI decision is augmented with confidence levels to enable the human to take an informed decision based on the limitations that are expressed through the confidence/uncertainty metrics.
5. **Resilience and reliability:** Critical infrastructures need to be resilient to failures and reliable in their operation. UQ contributes to this by ensuring that AI systems are not only accurate but also aware of their limitations. This awareness can lead to more robust designs and operational strategies that account for potential uncertainties.

In summary, UQ bridges the gap between human judgment and AI capabilities, ensuring that decisions made within critical infrastructures are not only data-driven but also cognizant of the inherent uncertainties. This leads to more resilient, reliable, and safe operational outcomes. UQ is receiving attention from standardization bodies, e.g., the German Deutsches Institut für Normung (DIN) recently developed general guidance and requirements for the development and use of methods for quantifying uncertainty in ML, DIN SPEC 92005<sup>19</sup>, where a potential follow-up international standard regarding UQ is currently under consideration by NA 043-01-42 GA of DIN.

### 3.2.2.2 ALGORITHMIC ASPECTS OF AI-BASED MODELS

#### 3.2.2.2.1 KNOWLEDGE-ASSISTED AI

Knowledge-assisted AI refers to an approach to AI that makes use of pre-existing knowledge, often in addition to data-driven elements. This thus often concerns hybrid approaches that combine learning

---

<sup>19</sup>

Can be downloaded (in English) from: <https://www.dinmedia.de/en/technical-rule/din-spec-92005/376619718>

elements with human knowledge and cover approaches referred to as ‘neuro-symbolic’ (Van Harmelen et al., 2019), ‘hybrid,’ or ‘informed’ (Von Rueden et al., 2021) in the literature. This knowledge can come from various sources, e.g., an existing heuristic or factual knowledge such as symbolic rules or a physics equation. Knowledge-driven elements are strong at helping an AI approach generalize, especially where little data is available. On the other hand, the provided knowledge might not cover all possible scenarios, and data-driven elements can exploit training data to cover such gaps.

The essential properties of knowledge-assisted methods are that they (1) can learn from data and (2) can take prior knowledge into account. Some authors consider that such knowledge should come from an independent source and be given by formal representations (Von Rueden et al., 2021). However, from a broader view, implicit representations of domain knowledge (such as an existing procedure or heuristic) can also be considered. Key elements by which methods for knowledge-assisted AI can be classified include the source of knowledge, the representation of knowledge, and the integration of knowledge (Von Rueden et al., 2021).

The focus in many ‘knowledge-assisted approaches’ is not to add functionality but to increase the performance of functional components of the system, especially where available data is scarce and/or is not representative of the entire domain of interest. Furthermore, elements driven by, for example, logic-based or procedural knowledge tend to be more understandable to human users. Thus, systems including such elements might generalize more systematically and be more transparent and auditable than systems fully driven by (deep) ML models.

Evaluation of knowledge-assisted approaches can take place in several ways. Primarily, the core task performance should be evaluated as measured by an objective function. In the context of AI4REALNET, it might specifically be interesting to consider performance both in ‘regular’ regimes as well as in ‘abnormal’ regimes, such as during a system outage. Since less training data is available in these abnormal regimes, it can be hypothesized that the effect of adding a knowledge component is more prominent in such regimes. Where a system has functional components that provide explanations and transparency, the effect of including knowledge assistance on those components should also be evaluated.

In the context of AI4REALNET, several sources of prior knowledge can be identified. For example, in the power networks and air traffic domains, the relevant physics equations are known well enough to be exploited in a possible solution. Furthermore, virtual simulators can be used as a (coarse) proxy for at least initial training when moving to physical systems. Such knowledge needs to be brought in an accessible format to allow developed approaches to be applied across domains.

#### 3.2.2.2 META-AWARENESS FOR AI ASSISTANTS

The combination of the human operator with the AI assistant forms the human AI team, as phrased by (Endsley, 2023b). This team presents complementary facets. While AI systems based on ML can process large amounts of data and learn complex patterns, the human operator is much more capable of managing unexpected (e.g., where there is no historical data available) and edge situations. Therefore, the operator is always in charge of the system. Moreover, in general, ML-based systems lack a model of causation that is essential to predict future events, simulate potential actions, or generalize to new situations. Data-driven decision-making in evolving situations requires not only the perception of the current state of the environment but also the understanding of what can possibly happen or is likely to happen in the near future (Endsley, 2023a). In addition to aleatoric uncertainty,

which represents the inherent randomness of the environment, epistemic uncertainty (Hüllermeier and Waegeman, 2021; Charpentier et al., 2022) is also a crucial component of the system.

An AI-based decision system should go beyond the current pattern-recognition paradigm provided by ML systems and be able to address, among others, the following requisites:

- Learn quickly with as few episodes of failure as possible (Charpentier et al., 2022).
- Flag anomalous environment states when it does not know what action to take or suggest (Charpentier et al., 2022).
- Perform well in untrained situations (i.e., with reduced epistemic uncertainty) and manage aleatoric uncertainty as well as noisy data. As discussed in (Tomsett et al., 2020), information about uncertainty (uncertainty-awareness) can lead to improved trust calibration from humans in the AI model's output in high-stakes decisions. The uncertainty can be presented either by probabilistic indices (e.g., standard deviation, inter-quantile range) or non-probabilistic representations (e.g., confidence level).
- Keep human operators informed of important changes in the managed system and external information without distracting them from their core tasks.
- Provide timely feedback on performance and guidance on correcting team errors.

Endsley introduced the concept of meaningful control (Endsley, 2023b), emphasizing the need for AI-based systems to have a level of meta-awareness. This awareness enables them to recognize situations that exceed their capabilities and prompt them to seek human assistance. Consequently, effective mechanisms for transferring control to humans are essential. This requirement of meta-awareness can be found in the operation of critical infrastructures where AI assistants can be used to aid human manual actions. For instance, in the reinforcement learning competition described in (Marot et al., 2022a), one goal was to evaluate if an AI agent has the ability to send alarms to the operator ahead of time when the proposed actions are of low confidence and avoid a human out-of-the-loop scenario. On the other hand, the issue of over-alarms was a risk to positive human-agent interaction, and thus, an attention budget was considered. This framework was built to have high levels of credibility, reliability, and intimacy.

Following these concepts and requirements, the AI4REALNET meta-awareness concept considers the following phases:

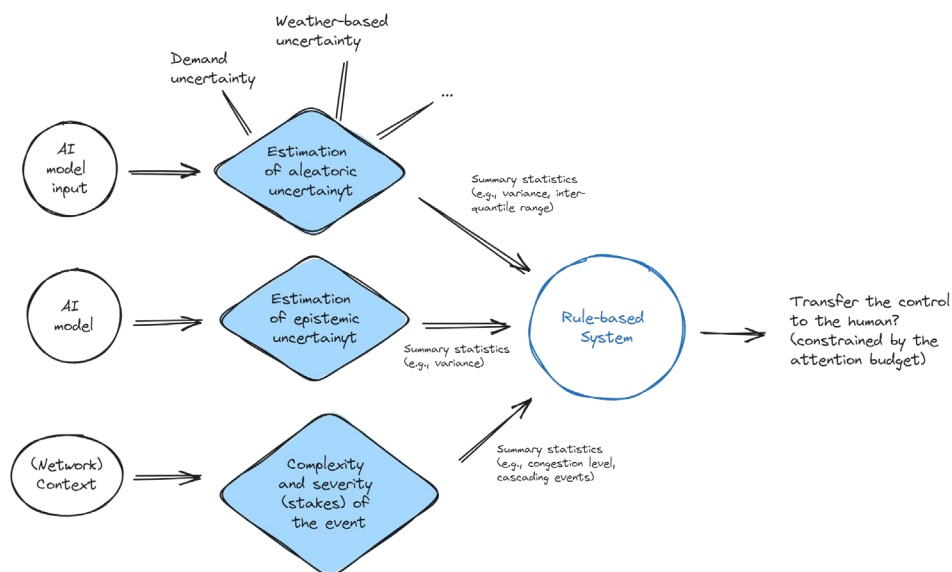
The AI-based system **observes and monitors** the managed infrastructure (e.g., power network). Developing this functionality of the system will require incorporating domain expertise, software engineering, and AI/ML and data science expertise to implement automatic data and information extraction systems that will be used to refine and calibrate the data-driven models. The capacity to derive and represent contextual information (e.g., knowledge graphs, structural causal model) about the operating context is fundamental for the learning and awareness properties, as well as for human understanding of recommendations and performance (Palminteri and Lebreton, 2021). This will allow AI systems to capture knowledge about themselves and the environment.

**Anticipation and alert** will be possible due to the predictive models refined from the system observation, as well as uncertainty quantification. This will allow us to forecast the workload of the managed system and the detection of anomalies and unexpected events. The combination of these

elements will allow us to predict the impact of the changes in the system environment and to detect and predict novel problems at different time scales during the managing operation period.

In this phase, both *aleatoric* and *epistemic* uncertainty should be considered to alert the human operator about situations where the sampling efficiency during the model training or the generalization performance is low (i.e., the model is experiencing “unknown states”) or situations with high stochasticity (e.g., high uncertainty in weather forecasting) that lead to low confidence in the recommendations/decisions. In this situation, the option should be to transfer the control to the human and provide all the information necessary for the human to decide (Nylin et al., 2022). The complexity of an environment’s operating conditions and events can serve as exogenous information (Campos et al., 2024), helping the AI system become aware of its own capabilities and enhancing its ability to provide accurate recommendations.

Figure 21 illustrates a prototype of a deferral mechanism that, following the nomenclature in (Bondi et al., 2022), learns to defer decision-making from the AI model to a human. This mechanism considers *aleatoric* and *epistemic* uncertainty, as well as the network context, and the rule-based system can also include a constraint related to the deferral rate (i.e., an acceptable level of human effort or the attention budget). This can be evaluated in real-time (i.e., for the current operating scenario) or predicted for the next lead-time where *aleatoric* uncertainty needs to be considered in the model.



**FIGURE 21 – PROTOTYPE SCHEMATIC OF A DEFERRAL MECHANISM THAT LEARNS TO DEFER DECISION-MAKING FROM THE AI MODEL TO A HUMAN**

### 3.2.2.2.3 HUMAN-AI CO-LEARNING

Work in the field of what this project refers to as “Co-Learning” can be found under many aliases, commonly a combination of “Human-AI” or “Human-Machine” with a suffix indicating the collaborative nature – “Teaming” or “Collaboration” under the most common. With recent advances in AI capabilities, research into the design of human-AI teams has gained momentum. First, discussions of how automated systems and humans will interact state that such systems must be perceived as individual and independent agents (Woods, 1996) and that autonomous agents must adhere to the principles of human-human collaboration (Rich and Sidner, 1997). Within the context of AI4REALNET,



we strive to achieve human-AI co-learning, which differentiates itself from existing research in that it aims to achieve continual and mutual learning in the human-AI team. It views the system holistically with the goal of exploiting strengths while mitigating weaknesses, thereby achieving performances superior to that which the agents could achieve individually.

A preliminary design is proposed here for a co-learning AI agent based on a paper titled “Six Challenges for Human-AI Co-Learning” written by Van den Bosch et al. in 2019. The paper provides a detailed description of co-learning and a discussion of requirements and challenges, focusing primarily on the agent side. The authors propose six models that an agent must have and continually refine to achieve mutual learning in a human-AI team. More specifically, an agent requires taxonomy, team, self, “Theory-of-Mind,” and communication models to be capable of interacting with human agents in a manner that conforms with human cognition (Van den Bosch et al., 2019). An overview of the system described in the following section is given in Figure 22, where arrows display interaction and information flow. This concept is merely descriptive, providing only an overview of the functionalities such a co-learning-capable AI must have without concrete concepts for technical implementations.

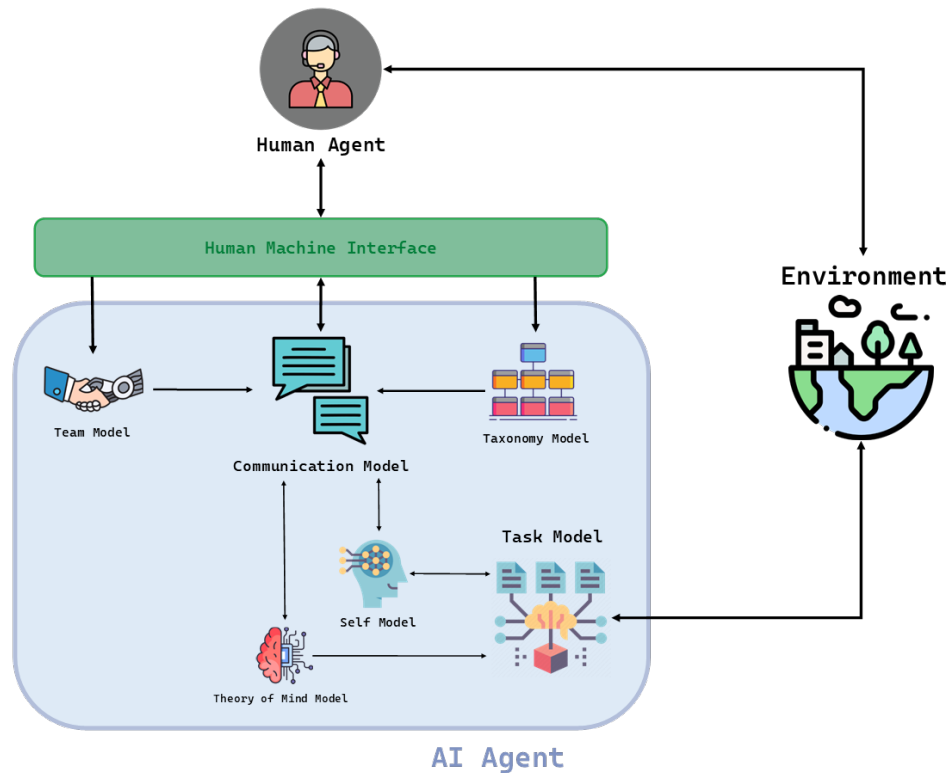


FIGURE 22 – DESCRIPTIVE SCHEMATIC OF A CO-LEARNING AI AGENT

For any human team to function, a common language and a shared understanding of team dynamics is required. In the proposed system by (Van den Bosch et al. 2019), the interaction between human and artificial agents is managed by the human agent via the **team model**, which defines work agreements, team organization, hierarchy, task distribution, and delegation. The **taxonomy model** manages the shared language pertaining to concepts and relations important for a common understanding of the task. With a common taxonomy and the agent’s place in the team defined, it can begin to solve tasks. To do so, a **task model** is required, which is comprised of knowledge about the

task and the relations between states, actions, and outcomes, including solution strategies and representation of state knowledge. Two models exist that describe the inner states of the team members: the **self-model**, depicting the inner state of the artificial agent, and the **“Theory of Mind”-model**, which covers knowledge about the inner state of other agents. Both models contain information about the goals, values, capabilities, resources, and intentions of the agents. The Theory-of-Mind-model differentiates itself from the self-model in that the information can be provided directly by the human agent or inferred by the artificial agent through behavioral observation. It also considers aspects of emotion and personality.

The knowledge of self and of others enables productive alignment and adaptation within the team, which occurs through the final model – the **communication model**, which is informed by the team and taxonomy model and exchanges information with the self- and Theory-of-Mind models. The sharing of information enables the AI agent to process human communication and send information using the defined vocabulary, under consideration of the human’s inner state, within the context of agreed-upon team dynamics while communicating its approach to the task as well as its own inner state. Communication of the inner state resulting from the self-model is of particular significance, given that it cannot be inferred from behavioral cues it would be in a human-human team (van den Bosch, 2019).

#### 3.2.2.2.4 MULTI-OBJECTIVE REINFORCEMENT LEARNING

##### Multi-objective learning in the context of Human Preference

A design strategy of multi-objective agents must begin with a discussion of the practicalities involved in training with multiple objectives and integrating human preferences. There are two steps involved in AI algorithm development: the training phase and the operation phase. In the training phase, a reward objective must be assumed to guide the training of the underlying parameters of the AI agent. In a multi-objective setting, this “total” reward  $R_{tot}$  is defined by a scalarization function  $U(R_1, R_2, \dots, R_n)$  of the individual objectives (Hayes et al., 2022). Assuming no human-machine interaction, this function  $U$  must be pre-defined at training time. An alternate solution would be to train an agent to determine the best solution under any mathematical combination of individual rewards, but this comes with the disadvantage of significantly increasing the complexity of the problem.

During the operation phase of the AI algorithm, for instance, in a control room setting, the algorithm may suggest one or many solutions to an emerging problem for the operator to choose from. These solutions are ranked by the total reward  $R_{tot}$ , and the individual rewards  $R_1 \dots R_n$  can also be calculated and presented individually to the human operator. When an experienced operator is presented with these solutions, they can select the solution that optimizes the total reward, or they may select a different solution based on intuition from their own experiences. This implies that the human operator’s preferences put a different weight on each individual reward, thus implying a different  $U$ -function (although this ideal  $U$  function may be unknowable and indeed inconsistent among operators).

Ideally, a “feedback” step would collect the disagreements between the operator and the AI algorithm’s predefined reward  $U$ -function. The goal of this feedback step is to align the preferences of the AI algorithm’s  $U$ -function with those of the human operator. Mathematically, this entails finding a  $U$ -function that replicates the order of the preferences chosen by the operator. Subsequently, an AI

algorithm retrained using this ideal  $U$ -function will order the solutions in a way that exactly matches the ordering of the expert human operator.

To achieve this, preference data of the human operator(s) would have to be collected over a wide range of scenarios. To improve the usefulness of this dataset, the operator could not simply choose the best action (out of, say, the 5 top choices that are suggested by the AI algorithm), but instead rank the actions from most to least desirable. Furthermore, cases in which the operator is indifferent to the choice between two actions should be noted as such to allow for the possibility that these actions lie on the “Pareto front” of the action space (Hayes et al., 2022).

The limitation of this approach is that it may be impossible to find the optimal  $U$ -function that aligns the AI algorithm with human preferences. Indeed, developing the dataset alone may be problematic. A more realistic approach would be to develop a  $U$ -function using heuristic methods, in cooperation with human experts, that align with human and AI goals.

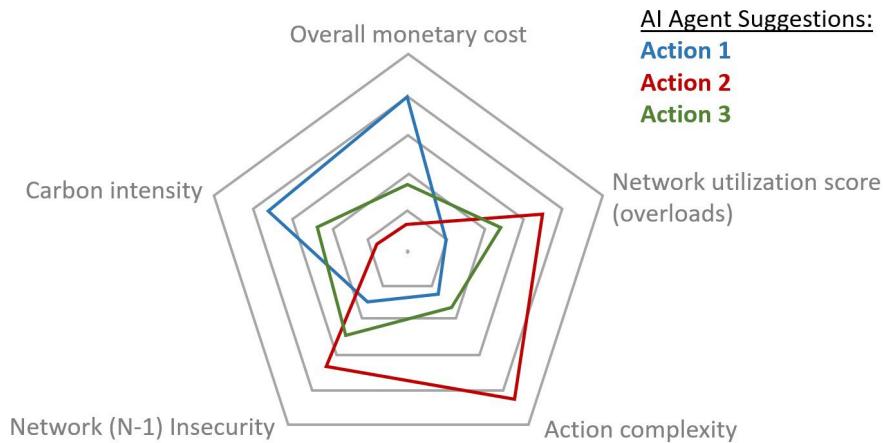
The previous description explains, in essence, the goals of the co-learning phase. To proceed to a fully automated setting, the learned  $U$ -function should be considered sufficiently aligned with human expert preferences up to a given tolerance. Retrospective analysis in the fully automated setting can ensure that the automated AI is performing properly and can provide additional training data for further refinements of the AI agents.

### **Roadmap for developing multi-objective AI Agents in co-learning scenarios**

Because the collection of human preference data is time-consuming and indeed requires an AI agent before proceeding, it is important to have a roadmap outlining the steps for achieving a multi-objective agent informed by human preferences. The outline below stresses that intermediate agents should be developed based on expert heuristics and, thus, should be highly effective agents even before the co-learning step has been reached. In addition, transparency and explainability regarding the breakdown of the total reward as a composite of sub-objectives are emphasized. The roadmap is as follows:

- 1) Identify the KPIs that should be converted into reward objectives in the context of RL.
- 2) In the absence of human interaction data, develop a heuristic model for obtaining  $U$ , the single-valued utility function.
- 3) Interim solution: provide visualizations of objective scores (e.g., a spider chart), which facilitates the objectives of explainability and transparency since it gives useful information about the chosen action, e.g., if a reward is a composite of multiple reward values, it is more transparent to give a breakdown of the individual reward scores so that the operator can make a more informed decision.
- 4) Develop a strategy for obtaining human preference data and record-keeping.
- 5) Implement a full human-feedback model to improve the consistency of human and AI decisions (utility function).

An example visualization of multi-objective reinforcement learning scores is depicted for the grid use case in Figure 23. In this scenario, each of the five objectives is intended to be minimized, and operators are presented with multiple plausible actions generated by the agent and their predicted scores on each metric. Recording the action preferred by the operator allows essential feedback for the improvement of the AI agent. Multiple instances of competing agents, trained to prefer different objectives, could also add variety to the set of possible actions.



**FIGURE 23 – EXAMPLE OF MULTI-OBJECTIVE VISUALIZATION**

### 3.2.2.3 HYPERVISION

Today’s supervision tooling is inherited from successive waves of IT implementation over the last decades: operator supervision over many screens and applications leaves the user the cognitive load to prioritize, organize, and link disparate displayed information and alarms before considering any decision or action.

More variable and complex infrastructure dynamics – driven, for example, by energy transition on electric transmission systems – tend to increase the complexity of tooling: in such a context, supervision becomes impractical, with numerous and complex information to process and non-integrated applications under heterogeneous formats. It contributes to the problem of information overload, which dilutes the operator’s attention. To be effective at continuous decision-making, it is often important to focus on the highest priority task at a time, using only the most relevant information. The sub-optimal design of human-machine interfaces and interactions has even been identified as a risk factor for human error in operations (Nachreiner et al., 2006).

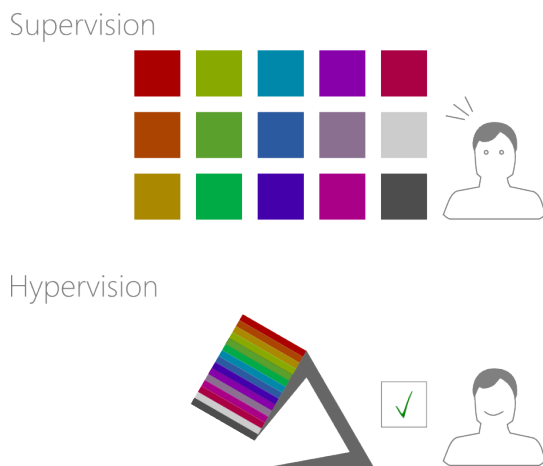


FIGURE 24 – FROM SUPERVISION TO HYPERVISION

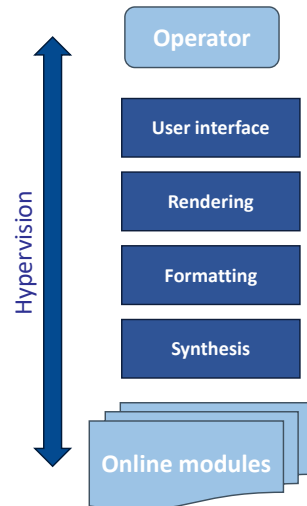


FIGURE 25 – HYPERVISION IMPLEMENTATION

In response to these challenges, the goal of Hypervision is to bring the right information at the right time to the right person while keeping track of user progress for each task (Marot et al., 2022b). From the operator’s point of view, Hypervision allows for synthesizing the necessary information and centralizing real-time business events in a single and unified interface supporting the decision-making process and prioritization of tasks to:

- Understand the operating context,
- Diagnose alerts,
- Choose the implementation of solutions.

Through a better prioritization and syncretization of events, Hypervision should allow for extending beyond real-time tasks and gaining a broader perspective to anticipate tasks to be completed or configured ahead of time thanks to forecast. By defining an adaptable trajectory, Hypervision shifts the focus from alarm monitoring to efficient task completion (see Figure 24).

Hypervision is implemented by four distinct layers (see Figure 25):

- **Synthesis:** This layer is connected to various existing online modules or tools and selects relevant information to use,
- **Formatting:** Select the most relevant way of presenting the information (text, table, graph, etc.),
- **Rendering:** Connect formatting and context (e.g., line overloads on a map)
- **User Interface:** Addresses synthesis and prioritization needs, human-machine dialogue, collaboration, and decision-making capitalization.

The Cockpit and Bidirectional Assistant (CAB) project aims to provide support in augmented decision-making for complex steering systems. The objective of the CAB project, launched in July 2020 for a period of four years, is the development and prototyping of a bi-directional virtual assistant – open in terms of industrial applications – in which it will be possible to evaluate the forms of exchange between the Human expert and an AI that continuously learns both from the information flows received and from the decisions made by human.

The interface of the CAB project (see Figure 26) is an example of a hypervision interface framework. It is composed of various panels:

- The context: It is the central panel of the interface and is where the environment and the context are visualized in real-time. The operator could monitor the different parts of the environment (power grid in this case) and the experience could be augmented using various offered functionalities (to zoom in on a specific area for example),
- Timeline: The timeline panel provided at the bottom part of the user interface allows us to monitor the time steps and also to keep a trace of various events in history. That enables the operator for example to go to previous time steps and to deepen the analysis,
- Alerts: The alerts panel located at the left part of the user interface shows the notifications about the context over time. These notifications could be related to the risks and events (for example the risk of overload on the power lines due to a disconnection),
- Recommendations: The recommendation panel is located at the right side of the user interface which provides recommendations using an AI-based agent. The operator has the choice to use this recommendation or not based on the expertise level and the complexity of the risks to be cleared.

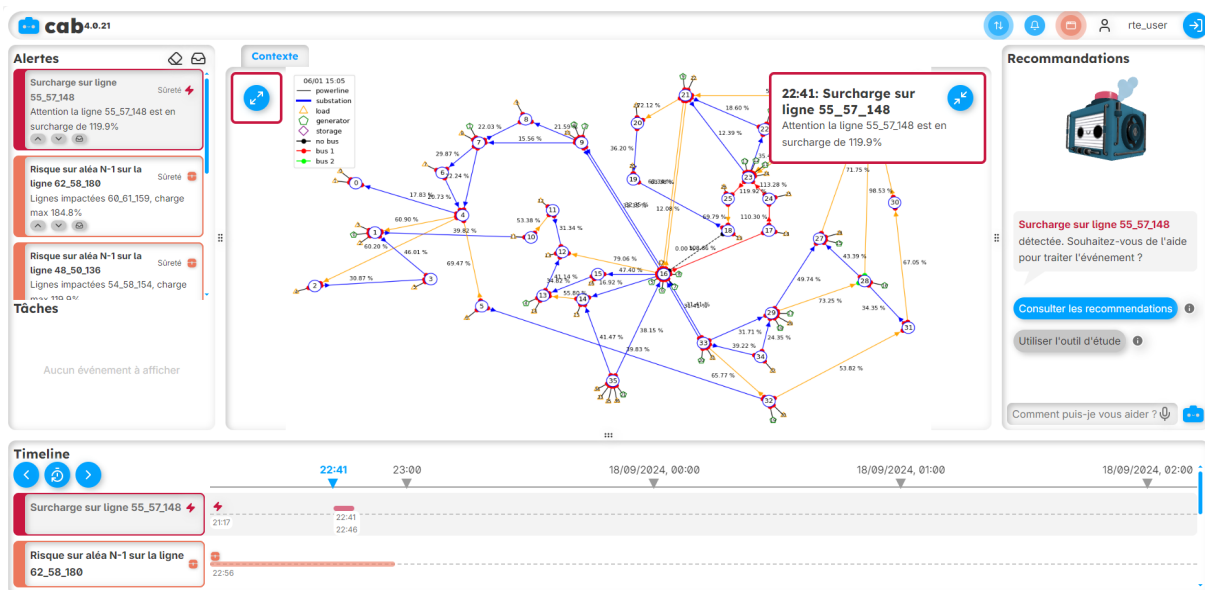


FIGURE 26 – EXAMPLE OF HYPERVISION INTERFACE (CAB PROJECT<sup>20</sup>)

The explanatory aspect of AI recommendations is central to the CAB project to give added value to the operator in their decision-making. The virtual assistant will be able to determine the profile of the operator and his cognitive workload level and adapt the information flows uploaded to the operator in order to manage a complex and/or atypical situation in the best conditions.

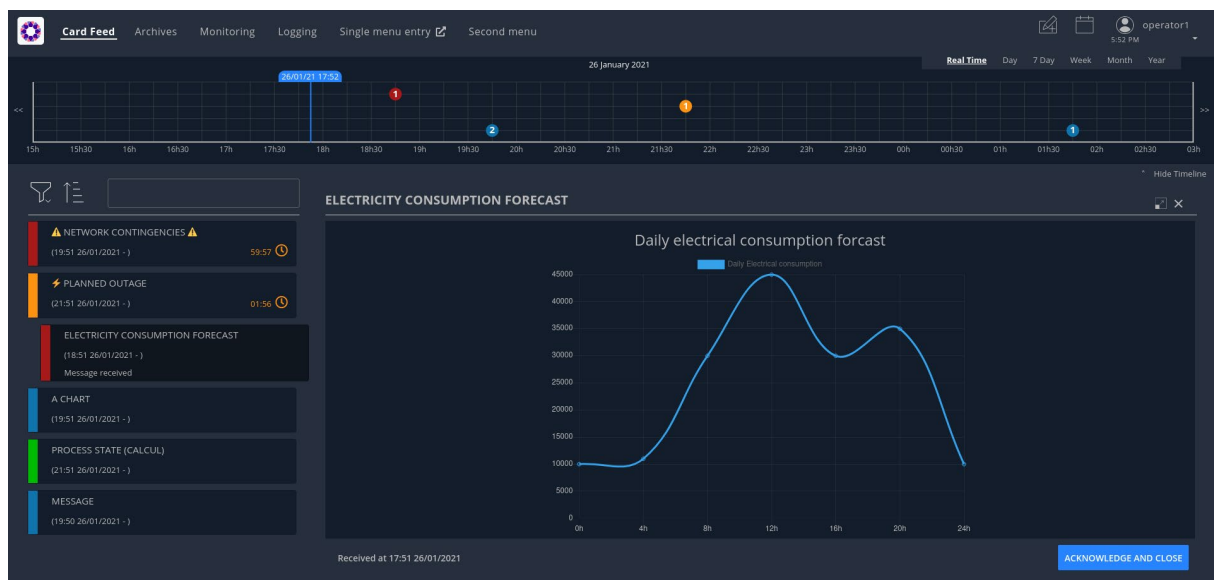
OperatorFabric<sup>21</sup> (see Figure 27) is another example of a Hypervision interface framework and interface that implements Rendering and User Interface layers. It regards the decision-making process as a task, represented by a digital card that is ordered by priority in a feed. When a card is selected,

<sup>20</sup> <https://www.irt-systemx.fr/en/projets/cab/>

<sup>21</sup> LFEnergy. (2019, Jul.). Operator fabric: a smart assistant for system operators. [Online]. Available: <https://opfab.github.io/>

the details of the card are displayed: information about the state of the process instance in the third-party application that published it, available actions, etc. The card lifecycle can be composed of the following steps:

- Automatic creation and notification to the operator based on forecasted alerts and contextual information with a preliminary diagnosis (or eventually with a procedure and configuration choices for execution),
- Tag by the operator as representing a certain type of problem and objective,
- Update (possibly automatically) and refinement as refreshed forecasts or new information come in, or manually edit by the operator (for example, in more unusual situations),
- Recommendations for actions can be also added within the card (or the operator can propose other ones),
- Selection of one given recommendation by the operator that will be considered as active,
- Sharing across operators (based on tags, groups, organizational entities, processes, etc.) allowing for effective coordination.



**FIGURE 27 - EXAMPLE OF HYPERVISION INTERFACE (OPERATORFABRIC)**

A card with versioning eventually represents the full decision-making process, which can be analyzed step-by-step or backward. All cards can be displayed on timeline or agenda views that complement the card feed views by allowing the operator to see briefly the status of processes for a given period.

As structured decision-making is applied to any field, Hypervision interface frameworks such as OperatorFabric can be used in the operation of different types of critical grid infrastructures; only the underlying information management remains domain-specific.

### 3.2.3 HUMAN-AI INTERACTION AND SYSTEM DESIGN

The system design level focuses on the description of the technical system while incorporating the perspectives of stakeholders of the system and the environment in which the system is intended to operate. To cope with the complex environment and tasks the system will operate in and execute, multiple viewpoints on the envisioned system are taken to derive both functional and non-functional

requirements and guide the development process. Specifically, the system design level described in the following focuses on three distinct views: The operational view captures characteristics of the intended use of the system in a real-world setting. A functional view analyzes the functions the system should be able to provide. Lastly, a logical architecture segments the system into logical units and a building block view outlining the technical structure of the system.

Further, the AI-based operation of critical infrastructure—i.e., employing a system with a substantial degree of automation operating in an environment where high-impact decisions must be taken in real-time—imposes particular quality demands. These demands concern functional suitability (the provided functions meet the need for a high degree), reliability (a specified performance level is maintained under specified conditions), and operability (understandable, learnable, and usable by and attractive to the user), as well as quality characteristics according to ISO 25010. Therefore, the system design level highlights and addresses the aspects of robustness, UQ, knowledge-assisted AI, human-AI co-learning, explainability, and *multi-objective RL* in the second part of this section.

While these initial views, definitions, and considerations outlined in this section play a crucial role in ensuring the system’s overall coherence and alignment between the different aspects during development, further work carried out during this project is intended to expand and refine the content represented in this section.

To describe the conceptual framework, we choose the systemic representation based on three levels of views (as illustrated in Figure 28): the operational view, the functional view, and the logical (or process) view.

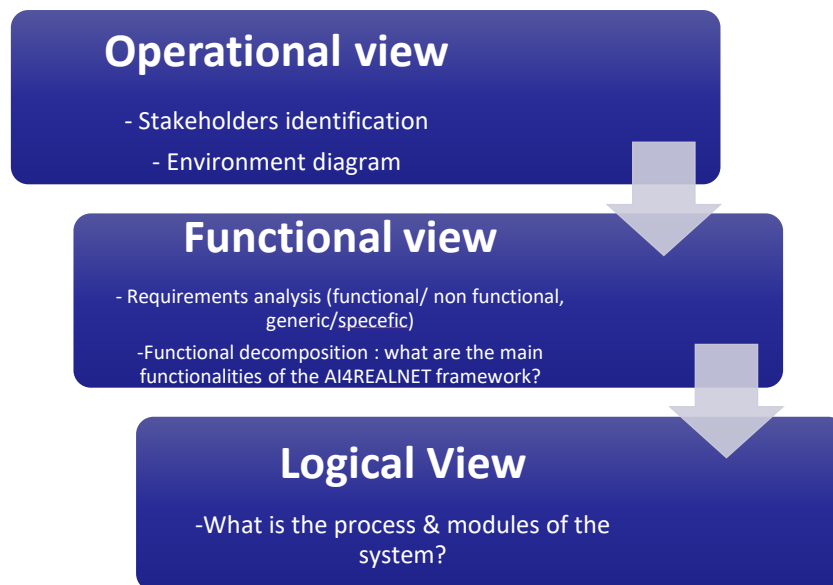


FIGURE 28 – GENERAL VIEW OF THE METHODOLOGY OF THE CONCEPTUAL FRAMEWORK

### 3.2.3.1 OPERATIONAL VIEW

The operational view in system design is a perspective that focuses on how the system will be used, operated, and maintained in its real-world environment. It addresses the operational aspects of the system, including the interactions between users, systems, and external entities.

#### 3.2.3.1.1 STAKEHOLDERS DIAGRAM



In this subsection, we describe all the stakeholders who interact with the system. A stakeholder is an external system that influences or interacts with the system to be designed. A stakeholder can be a human, an organization, or a technical system. It directly or indirectly impacts the system to be designed. It expresses a need or imposes constraints on the system.

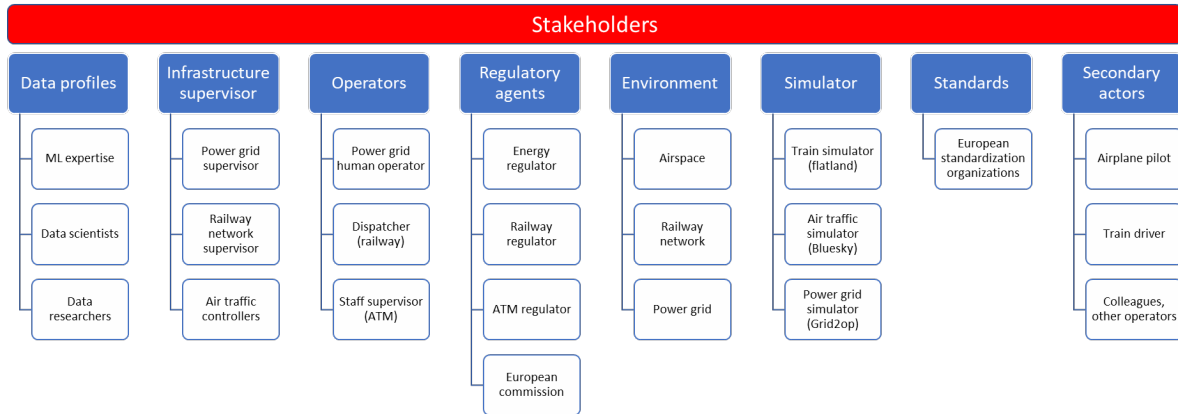


FIGURE 29 – STAKEHOLDERS DIAGRAM

As can be seen in Figure 29, the following stakeholders for AI-based critical infrastructure management were identified:

- **Data profiles:** manipulate the data at different levels of the conceptual framework to design AI-based algorithms and to provide meaningful decision support.
- **Infrastructure supervisor:** are individuals or entities responsible for overseeing the management, security, and operational efficiency of critical infrastructure systems. These supervisors ensure that the infrastructure operates smoothly, complies with regulations, and responds effectively to incidents or emergencies.
- **Operators:** They are key stakeholders in critical infrastructures, responsible for the day-to-day functioning and management of these essential systems. Their role is vital to ensuring the smooth, secure, and reliable operation of critical infrastructure services and is central in the decision-making process. They could be assisted or not by AI-based recommendation systems (e.g., power grid human operator, train dispatcher, staff supervisor in ATM).
- **Regulatory agents:** Regulatory agents in critical infrastructures are organizations and agencies responsible for overseeing and ensuring the safety, security, reliability, and compliance of essential systems and services. These agents establish regulations, guidelines, and standards to protect these infrastructures from various risks, including cyber threats, physical attacks, and natural disasters.
- **Environment:** It corresponds to the real-world environment in which the critical infrastructure operators are operating and should interact with other stakeholders or objects to perform their tasks (e.g., power grid, railway network, and airspace).
- **Simulator:** It corresponds to digital environments, allowing simulation of real-world environments. It allows operators to simulate the real-world context as well as the operation's impact and, hence, to perform more meaningful and reliable actions.
- **Standards:** European standardization organizations play a crucial role in system design by establishing guidelines, best practices, and standards that ensure interoperability, safety, security, quality, and efficiency.

- Secondary actors:** Secondary actors as stakeholders in critical infrastructures are those who are not directly involved in the operations but still play significant roles in supporting, influencing, or being affected by these infrastructures. These stakeholders can include regulatory bodies, suppliers, customers, emergency services, and others.

### 3.2.3.1.2 ENVIRONMENT DIAGRAM

An environment diagram in system design is a visual representation that illustrates the external entities, interactions, and contexts in which a system operates. This diagram helps define the system's boundaries and understand how it interfaces with external elements such as users, other systems, hardware, and environmental factors.

Figure 30 shows the interactions, with data flows, between the various stakeholders (human or systems) and the system as a black box. We can observe in this scheme that the environment provides the real-world context and data for the framework, which in turn are exploited by different stakeholders. As an example, the data profiles are used to train AI-based decision systems for human operators. Human operators and supervisors also interact with the environment and simulators. They try to analyze the impact of their actions on simulated environments before performing them in real-world environments. Furthermore, the operators interact also directly with the framework for the whole decision-support process. They could request assistance from the trained AI-based recommendation systems in different operational contexts. To ensure the security and reliability of the assistance, the regulatory agents analyze the decisions made by the framework to verify conformity with guidelines and standards.

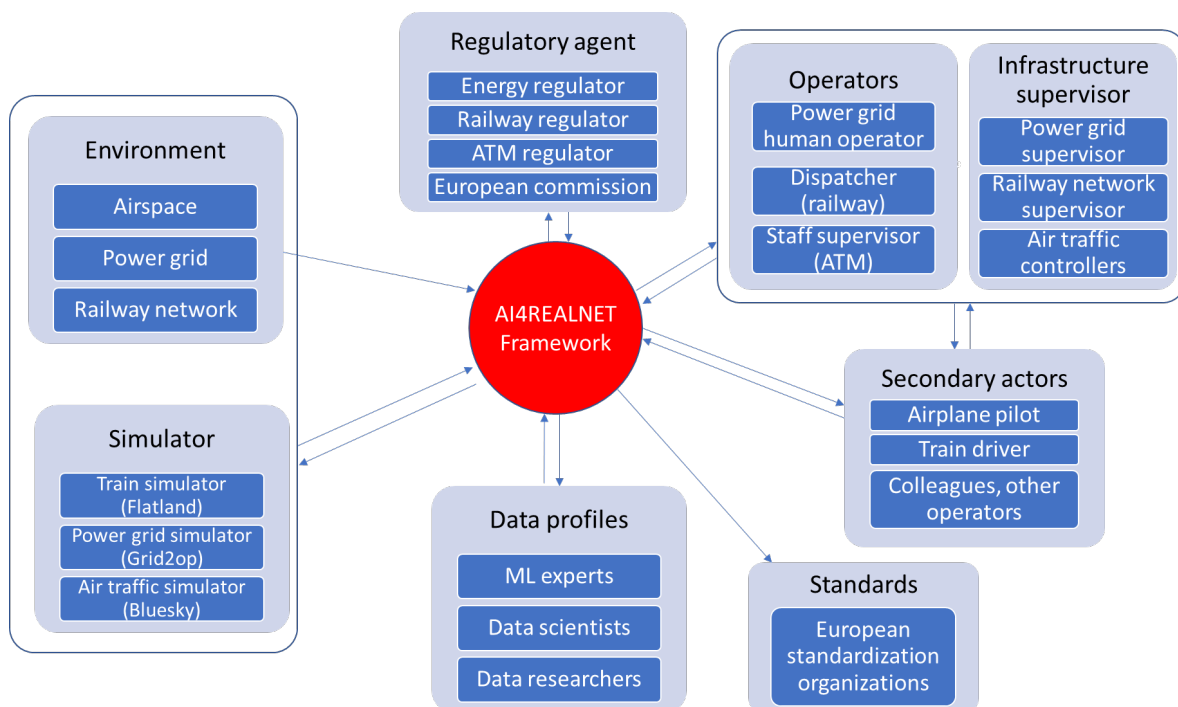


FIGURE 30 – ENVIRONMENT DIAGRAM

### 3.2.3.1.3 OPERATIONAL REQUIREMENTS

Operational requirements in system design specify the conditions under which a system must operate and the performance criteria it must meet during its usage. These requirements are essential for ensuring the system functions effectively in its intended environment and meets the needs of its users.

Operational requirements cover various aspects such as functionality, performance, reliability, security, and user interactions. These requirements are carefully addressed for each UC of the project and their corresponding scenarios in UC descriptions. We assign these requirements to functional vs. non-functional requirements and to generic vs. specific requirements (see Annex 2 for more details about the functional and non-functional requirements). This will help to design the functional view of the framework.

Based on the requirements described in the first part of this document, we assign Table 12, in front of each requirement, a pair of (F/N, G/S) will be used to designate functional (F) vs. non-functional (N) and generic (G) vs. specific (S) requirements. Generic requirements will be directly integrated into the functional analysis of the conceptual framework. In contrast, specific requirements (related to a specific use case) should be generalized to have a generic representation on the top of use cases.

Category	Power Grid	Railway	Air traffic
<b>Robustness</b>	<ul style="list-style-type: none"> <li>• Keep electrical grid security (F, S)</li> <li>• AI informs the human operator about its confidence in the output recommendation (self-awareness) (F, G)</li> <li>• Fault tolerance (F, G)</li> <li>• Reproducibility and traceability of recommendations for post-mortem analysis (F, G)</li> <li>• Adaptability to different operating conditions (F, G)</li> <li>• Do not increase cybersecurity risk (N, G)</li> <li>• Keep acceptable performance levels under natural or adversarial perturbations during operation (N, G)</li> <li>• Robustness to attacks targeting model space and reward function (N, G)</li> <li>• Detect changes in AI behaviour (F, G)</li> <li>• Adaptation to increased uncertainty (F, G)</li> <li>• Network change responsiveness (F, G)</li> <li>• Cognitive load and stress (N, G)</li> <li>• Reproducibility of recommendations for <i>post-mortem</i> analysis (N, G)</li> <li>• Increase technical robustness to missing or erroneous input data (F, G)</li> <li>•</li> </ul>	<ul style="list-style-type: none"> <li>• Reasonable recommendations in new situations (not seen during model training) (F, G)</li> <li>• Good performance in operating scenarios with high variability (N,G)</li> <li>• Retrospective quality control (N, G)</li> </ul>	<ul style="list-style-type: none"> <li>• System resilience to unexpected events (N, G)</li> <li>• Cyber and data security (N, G)</li> <li>• System’s reliable operation and decisions (N, G)</li> </ul>
<b>Efficiency</b>	<ul style="list-style-type: none"> <li>• Computational efficiency (N, G)</li> <li>• Relevance of the recommendations (N, G)</li> </ul>	<ul style="list-style-type: none"> <li>• Capacity to handle operating scenarios with high complexity (N, G)</li> <li>• Scalability (N, G)</li> </ul>	<ul style="list-style-type: none"> <li>• Capability to optimize resources and operations (F, S)</li> <li>• Scalability (N, G)</li> </ul>

Category	Power Grid	Railway	Air traffic
	<ul style="list-style-type: none"> <li>Scalability (N, G)</li> <li>Adequate training environment (N, G)</li> </ul>	<ul style="list-style-type: none"> <li>Generalization to different scenarios (F, G)</li> </ul>	
<b>Interpretability</b>	<ul style="list-style-type: none"> <li>Adaptability to different levels of interaction and human operator preferences (F, G)</li> <li>Action rating (F, G)</li> <li>Transparency during system training (F, G)</li> <li>Capacity to explain recommendation(s) to the human operator (and other stakeholders) (F, G)</li> <li>Adaptability to different levels of interaction and human operator preferences and experience (F, G)</li> </ul>	<ul style="list-style-type: none"> <li>Interpretability of suggestions (F, G)</li> </ul>	<ul style="list-style-type: none"> <li>Provide clear, understandable explanations for its decisions (F, G)</li> <li>Usability of the system from the human and other stakeholders' perspective (N, G)</li> </ul>
<b>Non-discrimination and fairness</b>	<ul style="list-style-type: none"> <li>Avoid creating or reinforcing unfair bias in the AI system (F, G)</li> <li>Regular monitoring of fairness (F, S)</li> </ul>	<ul style="list-style-type: none"> <li>Distribution of Delays (N, G)</li> </ul>	×
<b>Human Agency and Oversight</b>	<ul style="list-style-type: none"> <li>Additional training about AI for human operators (N, G)</li> <li>Mitigate addictive behavior from humans (N, G)</li> <li>Mitigate de-skilling in the human operators (N, G)</li> </ul>	×	×
<b>Regulatory and legal</b>	<ul style="list-style-type: none"> <li>Compliance with existing operational policies (N, G)</li> <li>European AI Act (F, G)</li> <li>Transparency to humans in terms of interaction with an AI system (N, G)</li> </ul>	<ul style="list-style-type: none"> <li>Compliance with legal standards and regulations (N, G)</li> <li>RUOM Favouritism (N, S)</li> </ul>	<ul style="list-style-type: none"> <li>Compliance with legal standards and regulations (N, G)</li> </ul>
<b>Data governance</b>	<ul style="list-style-type: none"> <li>Processing of human operator data (N, G)</li> </ul>	×	×
<b>Accountability</b>	<ul style="list-style-type: none"> <li>Allow audits for the AI recommendations and human operator actions (N, S)</li> <li>Reporting of potential vulnerabilities, risks, or biases (F, G)</li> </ul>	×	×
<b>Other</b>	×	<ul style="list-style-type: none"> <li>Maintainability (N, G)</li> <li>Environmental Sustainability (N, G)</li> </ul>	<ul style="list-style-type: none"> <li>Maintainability (N, G)</li> <li>Environmental Sustainability (N, G)</li> </ul>

TABLE 12 – CATEGORIES FOR THE THREE DOMAINS

## 3.2.3.1.4 HUMAN-IN-THE-LOOP AND OVERSIGHT REQUIREMENTS

In addition to the operational requirements intrinsic to the domain presented earlier, we have identified an additional set of functional requirements to enable human-in-the-loop decision-making under risk and uncertainty. To follow, we list in Table 13 the identified requirements by presenting a short description and a categorization based on the part of the system and the actor that must adhere to these requirements. The identified actors can be divided into *Operator*, the human interacting with the system; *Agent*, the AI-powered side of the decision-making process; and *environment*, either the true one or the cloned copy accessed by the agent for forecasting and assessment.

Requirements			
Categories ID	Category name for requirements	Category description	Actor
R-01	Alarm Triggering Human	An operator can trigger the alarm to interrupt an execution and step into the decision-making.	Operator
R-02	Inspect Status	An operator can inspect the system's undergoing situation to observe what has happened and what caused their intervention.	Operator
R-03	Provide Action	An operator can take action if the suggestions given by the agent are not exhaustive.	Operator
R-04	Inspect Remedial Plan	An operator can access and inspect a remedial plan proposed by the agent to see the sequence of actions autonomously scheduled by the agent.	Operator
R-05	Inspect Feedback	An operator can analyze the feedback provided by the autonomous agent to decide whether a recommendation should be followed.	Operator
R-06	Ask for Additional Recommendation	An operator can ask the agent to provide additional suggestions in case none of the remedial plans satisfies their expectations.	Operator
R-07	Provide Constraint	An operator can limit the freedom of an autonomous agent by providing temporary constraints to get remedial plans that respect the given indication. For example, in the power grid domain, an operator may force a certain substation to be switched. Consequently, the remedial plans provided by the AI must switch the given substation.	Operator
R-08	Recommendation Selection	An operator can decide to follow a remedial plan provided by the agent.	Operator
R-09	Estimate Epistemic Uncertainty	The agent can estimate the epistemic uncertainty of its decision model to establish its level of confidence within an observed state.	Agent
R-10	Alarm Triggering Agent	The agent can raise an alarm to draw human attention. Consequently, the human operator will need to provide some input.	Agent
R-11	Simulate Remedial Plan	An agent can interact with a copy of an environment to simulate a remedial plan and/or provide feedback.	Agent
R-12	Adapt Recommendation I	The agent can adapt its recommendation based on constraints given by a human operator.	Agent
R-13	Adapt Recommendation II	The agent can adapt its recommendation based on the human mental status to prevent additional stress on the human operator.	Agent
R-14	Handle Operator Plans	The agent can process a plan given by an operator to roll out an action or a sequence of actions and provide feedback.	Agent
R-15	Validate External Plans	The agent can validate and simulate an external remedial plan provided by an operator	Agent

Requirements			
Categories ID	Category name for requirements	Category description	Actor
R-16	Provide Recommendation	The agent can provide up to N multiple remedial plans. Each remedial plan consists of a sequence of actions to be taken to address a certain situation.	Agent
R-17	Provide Visual Human Readable State	An environment should provide a graphical depiction of the undergoing situation to allow human intervention.	Environment
R-18	Provide Text-based State	An environment should provide additional context in a text-based manner to extend the graphical depiction and allow the human to take informed actions.	Environment
R-19	Estimate Aleatoric Uncertainty	An environment should support the estimation for aleatoric uncertainty derived by an external source, such as weather conditions.	Environment
R-20	Clone	An environment should provide a method to clone and synch a simulated copy from the true environment instance. Cloning allows an AI agent to rollout actions from the current state and then provide feedback to an operator.	Environment
R-21	Action Conversion	An environment should support the bi-directional conversion of actions. From human-readable action to agent action and vice-versa.	
R-22	Raise Alarm	The system should have an alarm accessible by a different range of actors to enable a human to intervene.	Environment
R-23	Communicate Alarm Context	On the triggering, the system should provide extensive information about the causes that triggered the alarm	Other
R-24	Pause Interaction Loop	The autonomous Agent-Environment interaction loop must allow for being interrupted and paused when additional external input is required.	Other

**TABLE 13 – HUMAN-IN-THE-LOOP AND OVERSIGHT REQUIREMENTS**

Along with the UC-specific requirements, these requirements will contribute towards the design of the functional view of the framework by supporting the design of the human-AI interaction loop.

#### 3.2.3.1.5 OPERATIONAL USE CASES DIAGRAM

This stage enables us to understand how the system will be used and interact with stakeholders and an operational UC diagram (see Figure 31 below) is used to highlight the added value of the system to be developed.

We have identified eight operational UCs: From this system, an operator expects to receive contextual information (from his external environment and from the system being controlled), as well as a summary of current events. This information, which can be very numerous, needs to be sorted so that it can be communicated at the right time. The novice operator also expects to be assisted in his decision-making by considering both the external context (the environment and the system being piloted) and the internal context. An expert operator can contribute to the assistant's continuous learning (enriching its knowledge base). To distinguish the two cases, the operator can select their interaction mode, choosing between human in full control, co-learning or human as a supervisor with AI making automatic actions/decisions.

Finally, the operator should be able to communicate with external stakeholders to send or receive complementary information.

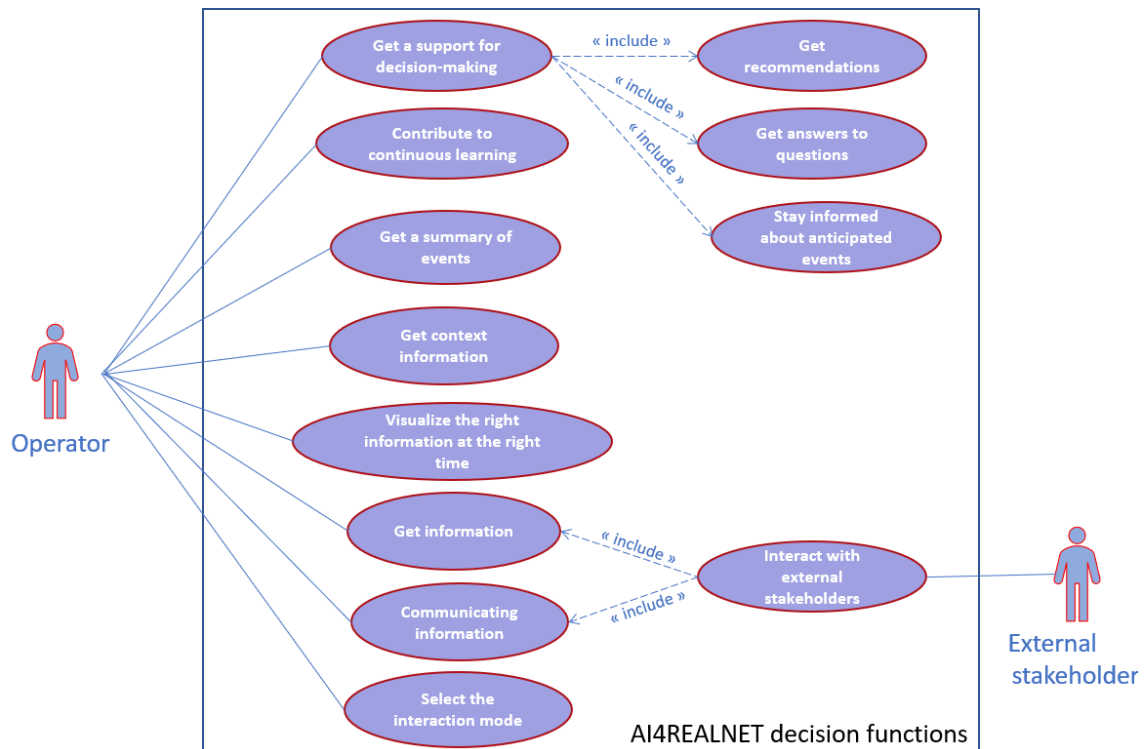


FIGURE 31 – OPERATIONAL USE CASES DIAGRAM

### 3.2.3.1.6 ABSTRACT BASE USER STORY

The use cases describe the interactions between the user and the AI-based decision system (see Annex 2 for detailed descriptions). Based on these descriptions and the general context of the UCs, we derived user stories during a cross-domain workshop which allowed us to identify commonalities between the use cases (see Annex 4). Further, this process enabled the distillation of an abstract base user story that can guide (together with the other parts of this framework) the development of the AI-decision system, especially concerning the human-machine interface, and enables to a certain degree the development of components and the design of interaction patterns applicable to all use cases across domains. The base story has three manifestations dependent on the time horizon: *prepare* for foreseen events (planning), *prevent* predicted events (near real-time), and *correct* events that happened (real-time).

All manifestations of the base story follow the same pattern: The story happens within a *context*, in which a *trigger* results in a series of *three actions*. First, a situation is observed (*context*), either through real-time monitoring (detecting differences between what is happening and what is planned), simulating what might happen (simulating potential futures), or by determining foreseen potential events. If a deviation is detected (real-time) or a potential deviation (planning and near real-time) is identified (trigger) by either the user or the AI system (depending on the UC), measures for the given situation are explored in all three manifestations. After exploring potential measures, either the human or AI system (depending on the interaction mode) is chosen to address the given situation. Finally, the story concludes with one of two actions that describe an intervention: either the chosen measure is implemented to correct the non-nominal state or to prevent a situation from happening,

or the measure is included in the operational plan in case the predicted or foreseen event will occur. Figure 32 depicts the abstract base user story for all three manifestations.

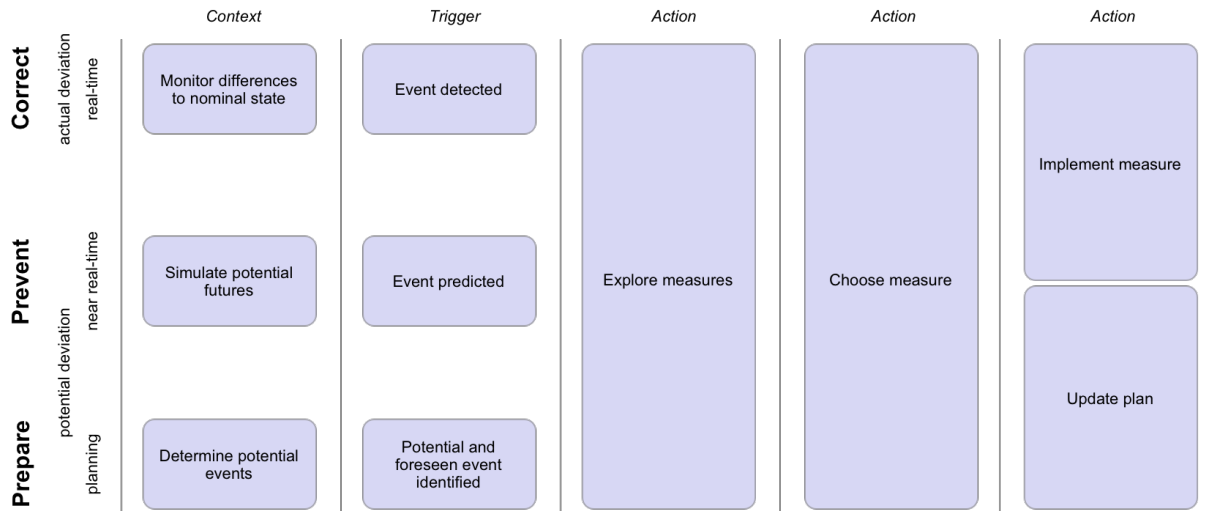


FIGURE 32 – ABSTRACT BASE USER STORY

### 3.2.3.2 FUNCTIONAL VIEW

This subsection is dedicated to defining the functions of the system, as well as their hierarchical decomposition. At this stage of the project, we have identified eight main functions for the AI4REALNET conceptual framework. Each function is further refined into several sub-functions. We present these functions using the functional decomposition diagram (see Figure 33). Main functions are represented by blue boxes and sub-functions with white boxes.

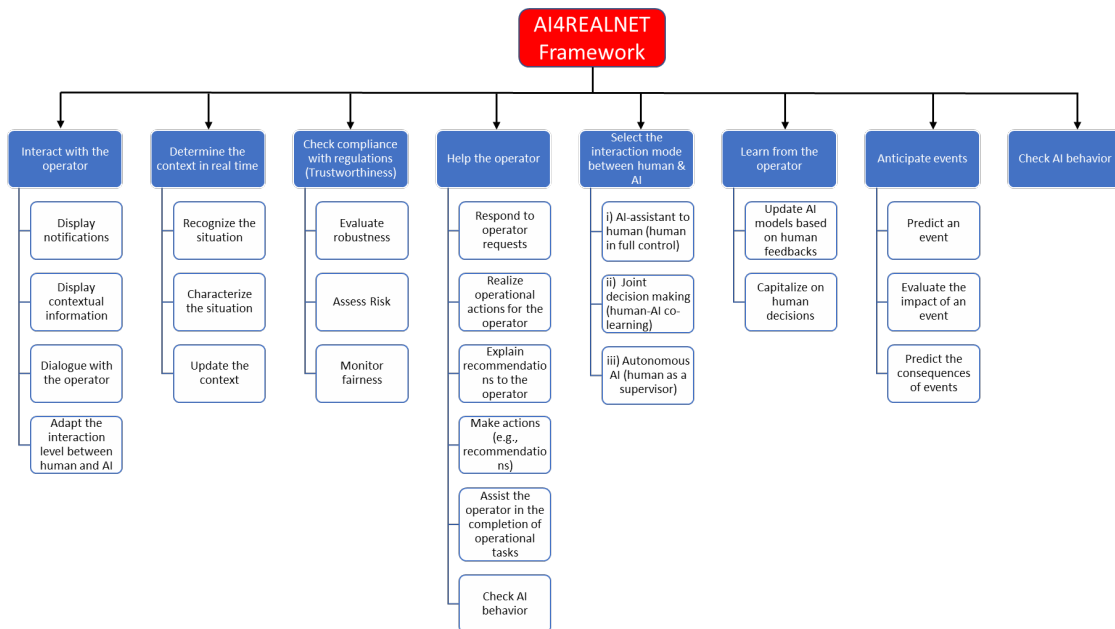


FIGURE 33 – FUNCTIONAL DECOMPOSITION

**Interact with the operator:** This function oversees all exchanges with the operator. It is the main interface between the operator and other system functions. It acquires commands and requests from the operator and then communicates responses to his needs (help, recommendations, explanations, information display, etc.). The needs may or may not be expressed by the operator. Additionally, **it**



communicates the assistant's requests/needs to the operator (e.g., help the assistant, provide exogenous information to complete the assistant's vision) and then receives the operator's responses to these needs/requests. This function also ensures that the right information is displayed in the right format at the right time (Hypervision concept, see Section 3.2.2.3).

**Determine the context in real-time:** This function oversees the collection and analysis of all data related to external (environment) or internal (controlled system) contexts. It is responsible for recognizing the operator's current actions, determining the external context (weather, for example), the internal context (the situation in which the operator finds themselves at a given time  $t$ , and the type and category of event they have to deal with at a given time  $t$ ). From these data, this function builds and saves the current context (over a given period of time). The assistant adapts to the context (external conditions, state of the controlled system, state of the operator, etc.). In this way, the output of this function is an important input for other functions, particularly the operator decision support function.

**Select the interaction mode between humans and AI: At the beginning of the mission, the operator should parameterize** the system by choosing its preference concerning its interaction with the AI. Different interaction modes can be available:

- AI-assistant to human (human in full control),
- Joint human-AI decision making (human-AI co-learning),
- Autonomous AI (human as a supervisor).

A default mode can be chosen.

**Help the operator:** This function is responsible for analyzing all operator events, commands, actions, etc., and providing the appropriate solutions/aids.

**Learn from the operator:** This function is responsible for extracting unacquired knowledge from the system and adding it to the knowledge/inference base. This knowledge can be provided by an operator, at the assistant's request, or on the operator's own initiative. It can also be based on the system's observations of the operator's gestures, actions, or behavior in the face of an unknown situation.

**Anticipate events:** This function is responsible for analyzing events received (situations, incidents, etc.) and analyzing historical and forecast data (weather, cultural events, etc.), then predicting future events and anticipating goals, actions to be taken, etc. It is also responsible for predicting the impact of current events; some failures or incidents can result in other incidents.

**Check compliance with regulations (Trustworthiness):** This function is responsible for assuring that the AI is always respecting and is compliant with regulations. For this, it continuously evaluates the Robustness, for example, by providing the confidence level and evaluating generalization capabilities. Also, this function could assess the risk and monitor fairness.

**Check AI behavior:** This function checks the AI behavior. This means that it continuously checks and calculates KPIs on AI outputs to be aware immediately of any change in AI behavior. The supervisor will be alerted if the AI malfunctions.

### 3.2.3.3 FUNCTIONAL INTERACTION DIAGRAM

Based on the environment diagram and the functional decomposition, the flow of data (functional objects) between the various functions will be determined, starting with the highest-level function down to the elementary functions or functions that interact directly with the stakeholders and then progressively identifying the missing internal functions (see Figure 34).

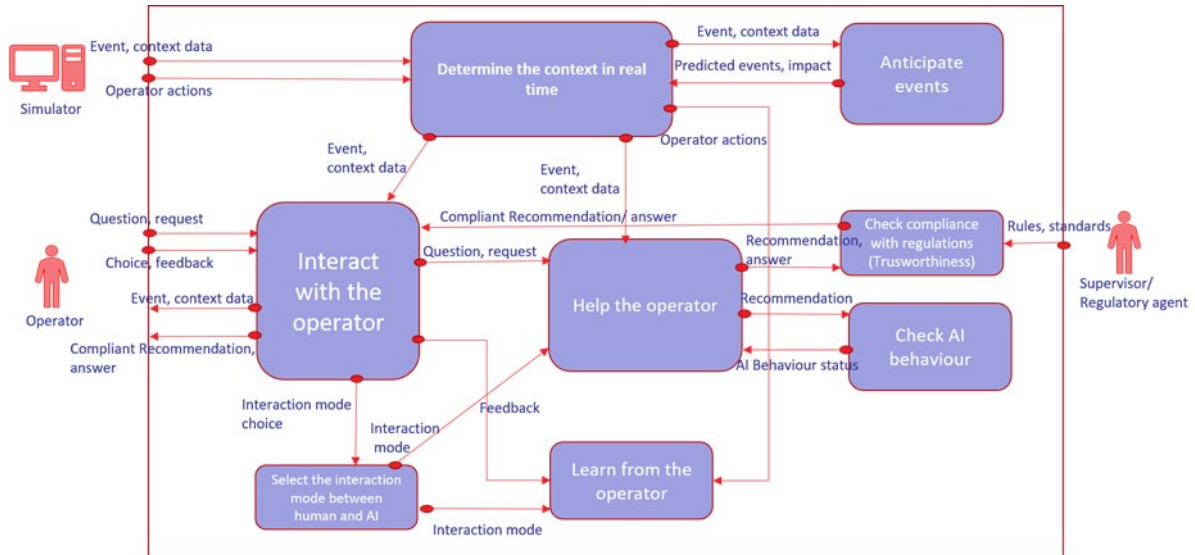


FIGURE 34 – FUNCTIONAL INTERACTION DIAGRAM

In this diagram, we consider the main eight functionalities presented in the previous section. We also add the interaction with the environment (operator, simulator, supervisor/ regulatory agent).

As can be seen in this diagram, the operator may interact directly with the platform to get some assistance, select the desired interaction mode (see the previous section for different modes), and finally, based on the selected interaction mode, give feedback concerning the provided recommendations which in turn will be stored as new knowledge. When assistance is possible through the platform and the selected mode of interaction, the recommendations are also verified to be compliant with regulations and reinforced with some KPIs on AI-based decisions.

### 3.2.3.4 LOGICAL ARCHITECTURE

The logical view in system engineering, also known as logical architecture, represents the abstract structure of a system. It focuses on the system’s functionality, decomposing it into logical components and their interactions without concern for the physical implementation. This view helps in understanding how the system meets its requirements and in identifying the relationships and dependencies among components. It is crucial to ensure that the system's design aligns with its intended purpose and facilitates communication among stakeholders by providing a clear, conceptual model of the system's functionality.

#### 3.2.3.4.1 PROCESS VIEW

The generic process is presented as an overview of the high-level interaction between different sub-systems which are decomposed to various functions. These functions were identified during the functional analysis. Within the conceptual framework and in the context of human-AI interaction, we

are interested in three different interaction levels: AI assistant to human (human in full control), joint human-AI decision-making (Human-AI co-learning), and autonomous AI (human as supervisor). In the following, we suggest three different logical views of the conceptual framework to represent the system components and interactions.

**Human in full control:** In the context of human-AI interaction, “human in full control” refers to scenarios where humans retain ultimate authority over decisions and actions influenced or assisted by AI systems. This concept emphasizes that while AI can provide insights, recommendations, or even perform tasks, the final decision-making power rests with humans. Overall, maintaining human control in AI interactions ensures that technology serves to augment human capabilities while safeguarding against unintended consequences or misuse.

The logical view corresponding to this mode of interaction is shown in the diagram of Figure 35. The environment in critical infrastructures is monitored in real-time using various sensors. It is represented in a specific context, which constitutes the observation space. We could define the decision boundary and characteristics (action space) based on the observed context. The digital environments provide us with a set of tools to simulate real scenarios, which in turn enables the assessment of the decision's impact before their application in a real-world context. When operating on the infrastructures, human operators should take some actions (decisions) to remedy the potential encountered problems. They could optionally take advantage of AI assistance to augment their capability at the decision-making step. The AI assistance is also accompanied with some explanations based on numerical indicators (decision support in the scheme) to guide human operators for selection of recommendations. Once a candidate's decision is made by the human operator, the regulatory agent can verify the trustworthiness of the decision through various KPIs.

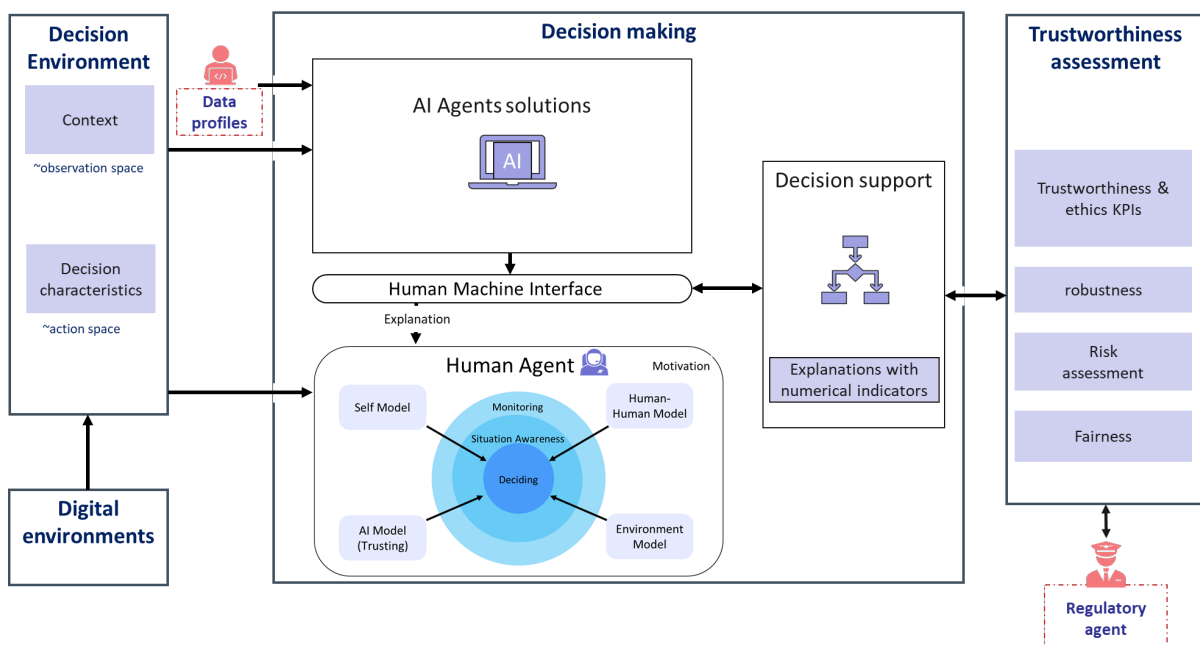


FIGURE 35 – LOGICAL ARCHITECTURE (HUMAN IN FULL CONTROL SCENARIO)

**Human-AI co-learning:** Human-AI co-learning in the context of critical infrastructure involves a synergistic partnership where humans and AI systems continuously learn from each other to enhance the efficiency, reliability, and resilience of essential services. This collaboration is crucial for managing

infrastructure such as power grids, water supply systems, transportation networks, and cybersecurity frameworks.

In this co-learning process (see Figure 36), AI systems can analyze vast amounts of data in real-time, identify patterns, and predict potential issues before they occur. For example, in a power grid, AI can monitor the network and detect anomalies that might indicate a fault. Human operators, on the other hand, bring contextual understanding and decision-making capabilities that AI lacks. They can interpret AI-generated insights within the broader context of socio-economic and environmental factors, make nuanced decisions, and adapt strategies as needed.

In co-learning, the human learning process is explicitly supported by AI to increase human decision-making skills. The overarching goal is to continuously improve human mental models about the environment, the AI, the self, and the cooperation with other people. AI can support these learning processes in different ways (e.g., by checking human assumptions or by mirroring his/her decision-making patterns). It is crucial that the collaboration between humans and AI is deliberately designed in such a way that it supports the human learning processes.

Moreover, humans can provide feedback to AI systems, refining their algorithms and improving their accuracy over time. This feedback loop ensures that AI systems are not static but evolve based on real-world experiences and expert knowledge. In critical infrastructure, this means that AI can help anticipate and mitigate risks more effectively, to improve human-AI joint decision-making.

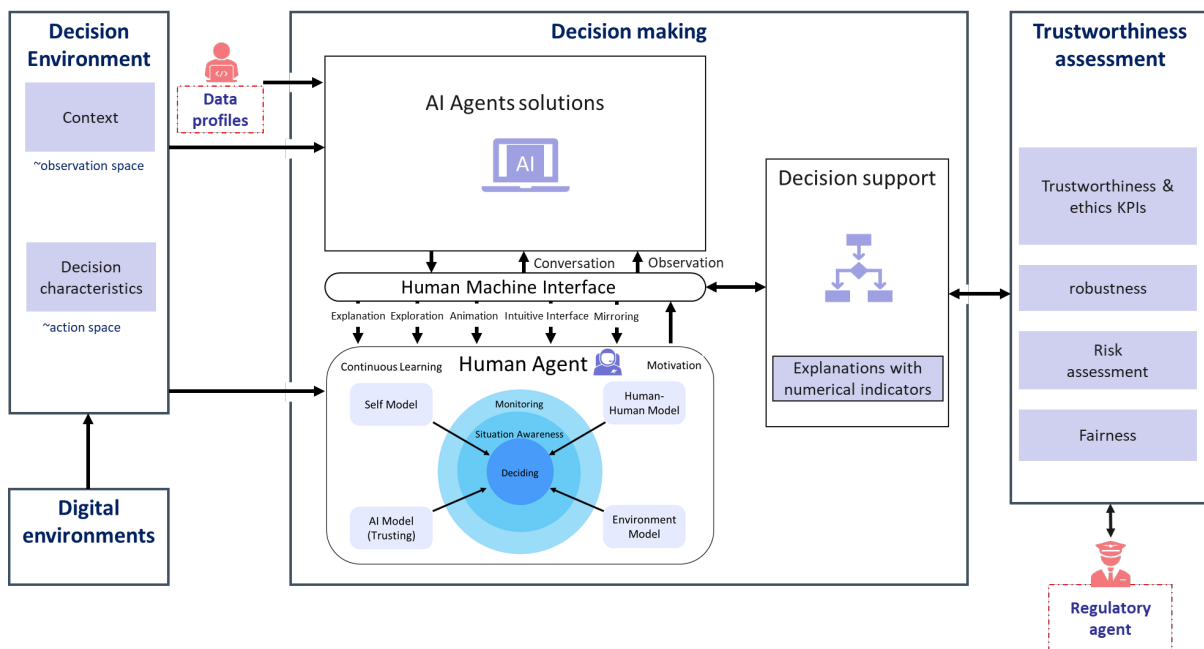


FIGURE 36 – LOGICAL ARCHITECTURE (HUMAN-AI CO-LEARNING SCENARIO)

**Human as supervisor:** Autonomous AI systems with human supervision (see Figure 37) in the context of critical infrastructure, which refers to AI technologies that operate independently to manage and control essential networks like the power grid, railway, air traffic sectors, and information and communication networks. These AI systems use advanced algorithms and ML to monitor, analyze, and make decisions to optimize performance, detect anomalies, and respond to emergencies.

However, given the high stakes and potential risks associated with critical infrastructure, human supervision remains crucial. This supervisory role involves overseeing the AI’s decisions, intervening in

complex or unforeseen situations, and ensuring that the AI operates within ethical and regulatory boundaries. Humans provide the necessary oversight to manage the AI’s limitations, address biases, and make judgment calls that require human intuition and experience. This is an extremely demanding task for humans and, therefore, requires appropriate automation transparency as well as targeted leverage points for interventions.

In summary, while autonomous AI can significantly enhance the efficiency and reliability of critical infrastructure, the human supervisor ensures safety, accountability, and compliance, creating a balanced and effective system.

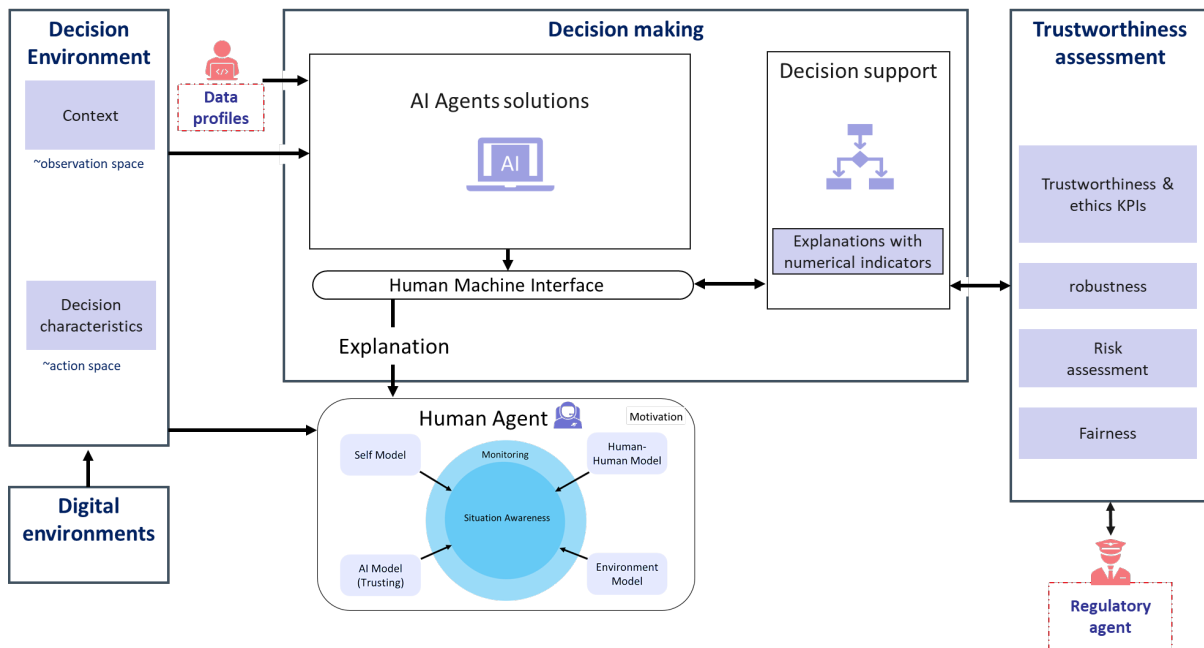


FIGURE 37 – LOGICAL ARCHITECTURE (AUTONOMOUS AI SCENARIO)

As we can see in these diagrams, the decision-making module is composed mainly of AI agent solutions and decision support. The human agent is considered as part of this module only in full-control and co-learning mode.

**Difference between AI agent solutions and decision-support:** The AI agent solutions integrate different AI models that generate recommendations. These generated recommendations are modeled according to each corresponding AI model and need supplementary processing to be displayed to humans in the interface. The decision support is responsible for processing the AI outcomes and generating human-friendly recommendations that the operator can understand. Also, the decision support can provide KPIs for each recommendation to help the operator compare and choose the more efficient recommendation.

3.2.3.4.2 BUILDING BLOCK VIEW

To strengthen the connections between the research questions in this project and increase the relevance of our findings for the development of integrated applications, we developed a high-level **conceptual prototype**: the AI4REALNET system. This prototype provides a practical framework to test and refine ideas, ensuring that the research outcomes are aligned with real-world needs. It will be refined throughout the project and can serve as early design requirements for future applications.

This system is described here via the building block view, which offers a hierarchical representation of the system from a technical perspective. Thereby, the system is decomposed into technical elements like modules, components, and frameworks, as well as the dependencies that collectively build the system. In addition, the building block view also shows interactions with users and neighboring systems. Figure 38 shows the scope, context, and high-level view of the AI4REALNET AI-based (conceptual) system, and the depicted blocks are described in the following.

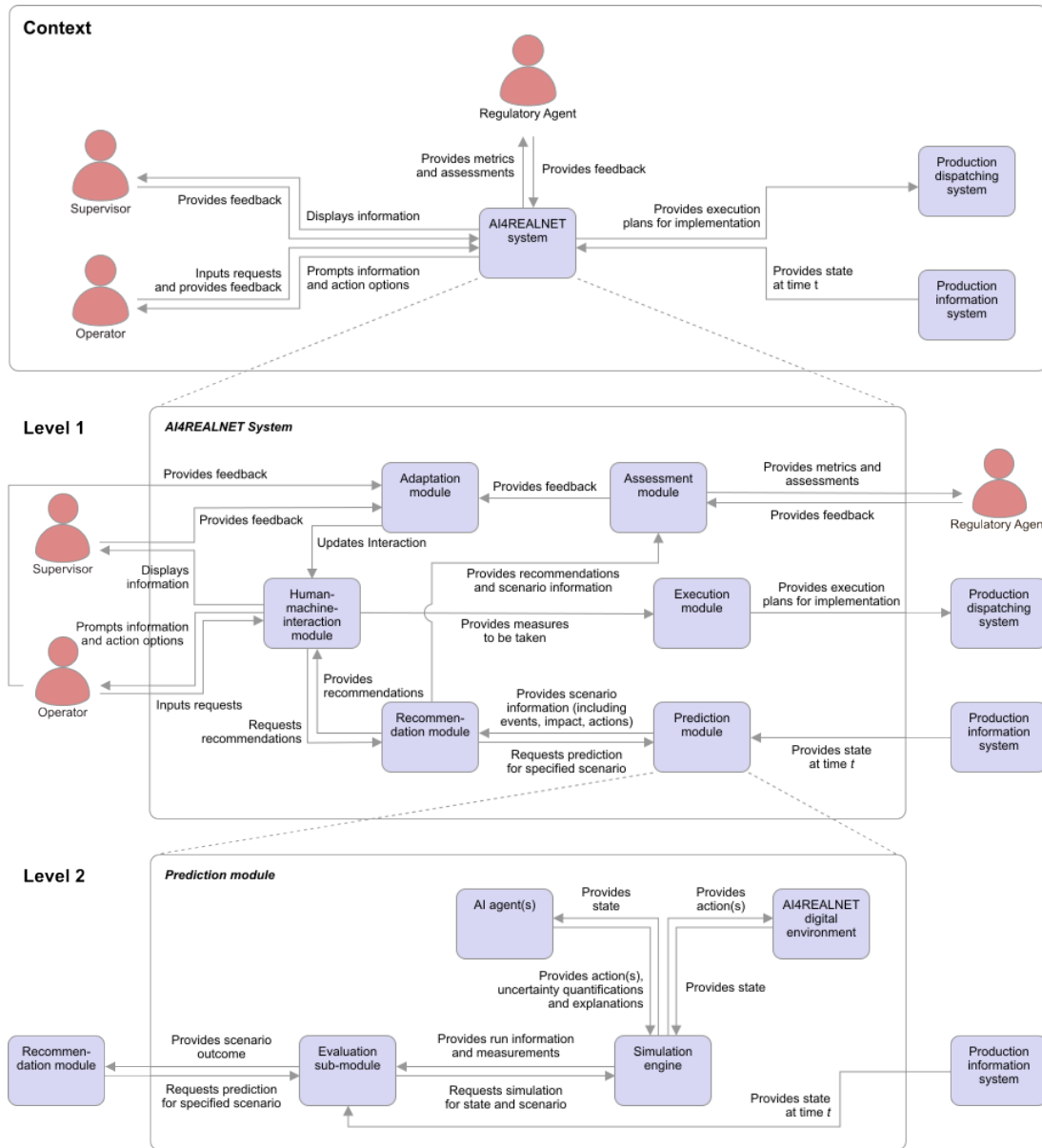


FIGURE 38 – HIERARCHICAL REPRESENTATION OF THE SYSTEMS’ BUILDING BLOCKS AND CONTEXT

### Scope and context

The AI4REALNET system is the center piece of the building block view and includes the AI parts aiding operations, the interfaces for human interactions as well as functions to directly assess its performance and to learn from feedback. The context in which the system operates includes neighboring systems to provide real-time operational information (production information system) and to implement decisions taken within the system in live operations (production dispatching system). Further, users,

such as operators, supervisors, and regulatory agents, are also part of the context and interact with the system.

### Level 1

The system is decomposed according to the nature of its function, e.g., functions that enable human-machine interaction are grouped into the Human-Machine-Interaction module:

#### *Human-Machine-Interaction module*

Functions and sub-functions: interact with the operator (display notifications, display contextual information (by implementing Hypervision concepts, see section 3.2.2.3), dialogue with the operator, respond to operator requests), help the operator (assist the operator in the completion of operational tasks), select the interaction mode between humans and AI.

This module should save capitalization data to a data store.

#### *Adaptation module*

Functions and sub-functions: determine the context in real time (recognize the situation, characterize the situation, update the context), interact with the operator (adapt the interaction level between humans and AI), learn from the operator (capitalize on human decisions, update AI models based on human feedback).

Capitalization on a human decision allows the module to update its training based on a data store containing all decisions, actions, etc.

#### *Prediction module*

Functions and sub-functions: anticipate events (predict an event, evaluate the impact of an event, predict the consequences of events).

This module should save capitalization data to a data store. The module gets the current state from the *digital environment* (which, in the production phase, could be the production information system).

#### *Recommendation module*

Functions and sub-functions: help the operator (make actions, explain recommendations to the operator).

This module should save capitalization data to a data store (including KPIs).

#### *Execution module*

Functions and sub-functions: help the operator (realize operational actions for the operator).

This module provides execution plans for implementation in the *digital environment* (which, in the production phase, could be the production dispatching system).

#### *Assessment module*

Functions and sub-functions: check compliance with regulations (evaluate robustness, assess risk, monitor fairness), check AI behavior.

### Level 2

At this early point, the structure of one module that plays an integral role in the early phase of the project is described in more detail.

#### *Prediction module*

The prediction module can be functionally decomposed into the *evaluation sub-module*, the simulation engine, the AI agent(s), and the AI4REALNET digital environment. The *evaluation sub-module* gets a state at a specified time  $t$  from the production information system and receives requests from the recommendation module. The *evaluation sub-module* is based on the current state of the digital environment (which, in the production phase, could be the production information system) and requests simulations from the simulation engine, which orchestrates the simulation run with the *digital environment* and the *AI agents* for specified states and scenarios given by the request. Eventually, the *evaluation sub-module* evaluates the simulation results to identify potential events, assess the impact of potential or occurred events, and identify potential consequences of an event and provides the outcomes together with other information like UQs and explanations to the *recommendation module*.

The *simulation engine* should save capitalization data to a data store (including KPIs).

### 3.3 EPISTEMOLOGICAL AND PHILOSOPHICAL FOUNDATIONS OF TRUSTWORTHY AI

This subsection investigates the epistemological and normative foundations of the notion of TAI and analyses the different components of risk and their application to AI with a particular focus on safety-critical systems. The goal is to lay the ground, from an epistemological and philosophical perspective, for a non-calculative approach to AI risk assessment. The starting point is the assessment list for TAI (ALTAI) elaborated by the high-level expert group appointed by the European Commission. The endpoint is a revised and improved ALTAI that focuses on key requirements for safety critical systems and takes into consideration, when needed, the three main components of risk (hazard, exposure, vulnerability). Overall, this part of the conceptual framework aims at devising a theoretical approach capable of dealing with risk and uncertainty that is difficult to quantify, suggesting that some problems must be addressed with methods that have a philosophical nature.

#### 3.3.1 THE EPISTEMOLOGICAL AND NORMATIVE GROUNDS OF THE NOTION OF TAI

The notion of TAI has been playing an increasingly central role in discussions on the responsible and ethically acceptable development and deployment of AI systems. Most notably, it provides the conceptual, philosophical, and ethical grounds for the effort the European Union has been making to provide an ethics-based regulation for the design and deployment of AI systems. Given the centrality of this notion for AI4REALNET, the analysis of its epistemological and normative grounds is a priority and complements the approach developed in section 3.2.1.22.4.

According to the Ethics Guidelines for Trustworthy AI of the European Commission, TAI has three components, which should be met throughout the system's entire life cycle: (1) it should be lawful, complying with all applicable laws and regulations (2) it should be ethical, ensuring adherence to



ethical principles and values and (3) it should be robust, both from a technical and social perspective. Despite its growing importance, the notion of TAI is not immune to criticism. First of all, there is no agreement on the determinants of trustworthiness in AI, namely, on what makes an AI system trustworthy. Moreover, it is not clear that some features that are typically deemed necessary for TAI are actually feasible for all AI systems. A prominent case involves explainability, which is systematically taken to be a fundamental ingredient of TAI and yet is hardly achievable in many systems.

Most importantly, however, there is an additional, foundational problem: from a conceptual point of view, it is unclear whether the very ascription of trustworthiness to AI systems could be a legitimate move (Nickel et al., 2010). Some authors have even argued that the “Trustworthy AI story is a marketing narrative invented by industry, a bedtime story for tomorrow’s customers. The underlying guiding idea of a “trustworthy AI” is, first and foremost, conceptual nonsense. [...] the Trustworthy AI narrative is, in reality, about developing future markets and using ethics debates as elegant public decorations for a large-scale investment strategy” (Metzinger, 2019).

Without getting into the details, the problem stems from the fact that standard accounts of trust and trustworthiness, which systematically take *interpersonal* trust and trustworthiness as models for these relations, typically assign a central role to the trustee’s interests, motivations, and moral obligations (Ryan, 2020). On these grounds, many voices called into question the ascription of trustworthiness to AI systems, which simply do not possess motivations and intentions and cannot adhere to moral obligations. Accordingly, the notion of TAI would be a categorical error, and its use would amount to some form of ethics-washing.

Building upon an awareness of these potential criticalities, the AI4REALNET project’s deployment of the notion of TAI starts from the acknowledgment that talk of trustworthiness in application to AI systems offers significant advantages. Most notably, it allows us to capture with a single notion two crucial dimensions of responsibly developed AI systems, namely reliability – *i.e.*, accuracy, and robustness – on the one hand, and ethical acceptability on the other. Among other things, this way of understanding trustworthiness in AI seems to ground the approach adopted in the European Ethics Guidelines for Trustworthy AI and the related ALTAI, that AI4REALNET takes as a starting point to identify the relevant risks involved in the use cases, develop specific requirements, and possibly provide tools for validating such requirements. In fact, among the requirements for TAI outlined in the guidelines, only the one of “technical robustness and safety” explicitly addresses aspects related to accuracy and robustness. The other requirements, instead, concern the ethical and societal impact of AI systems (accountability, human agency and oversight, privacy and data governance, transparency, societal and environmental well-being, diversity, non-discrimination, and fairness).

Even if the notion of TAI can provide significant advantages in keeping technical and ethical aspects together, the conceptual tenability problem remains to be solved. On the one hand, as a matter of fact, the notion of TAI encompasses an ineliminable ethical dimension. On the other hand, this dimension cannot be the same as interpersonal trust and trustworthiness, for this would require problematic attributions of paradigmatic human features to AI systems (again, having motivations and responding to moral obligations) to AI systems.

The approach adopted in the context of AI4REALNET involves the rejection of a widespread – and yet seldomly justified – methodological assumption in the philosophical literature on TAI, namely, the notions of trust and trustworthiness in AI should be *uncompromisingly* modeled on their interpersonal counterparts (Petrolo, G., Chiffi, & Schiaffonati). On the contrary, following (Zanotti et al., 2023), room is left for a conceptual distinction between trust and trustworthiness in human-human (H-H) *versus* human-AI interactions (H-AI).

As shown in Figure 39, a common conceptual core remains a distinctive feature of trustworthiness: just like in interpersonal relations, trust, and trustworthiness in human-AI interactions involve an aspect of reliability and encompass an ineliminable ethical component. The point is that when it comes to the way the ethical component is realized, H-H and H-AI trust differ. While trustworthy humans have the right interests, act upon goodwill, and adhere to moral obligations, TAI systems comply with specific ethical requirements. For instance, looking at AI4REALNET’s use cases, and in particular, at the systems involved in the use case *Sim2Real, transfer AI-assistant from simulation to real-world operation*, the design of the tool needs to be driven by the aim of avoiding human manipulation (e.g., misleading feedback, deliberately misusing the AI learning process).

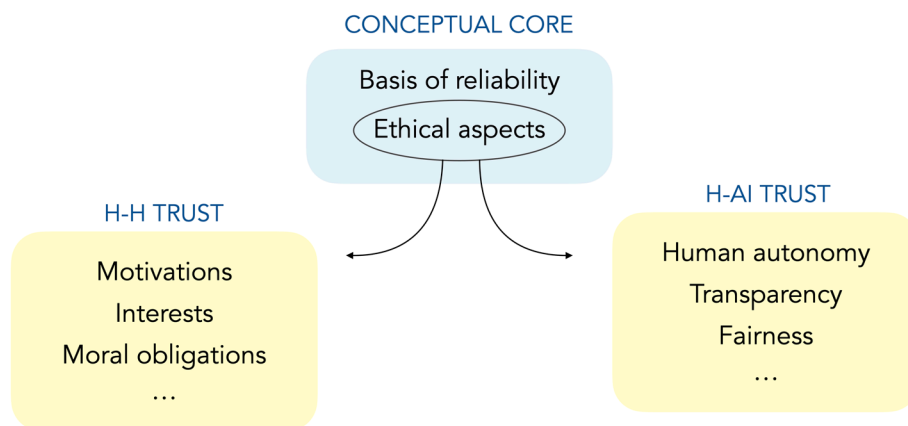


FIGURE 39 – H-H VERSUS H-AI TRUST

As it is acknowledged that the ethical dimensions of trust and trustworthiness in H-H and H-AI interactions are different, the conceptual error risk is averted, and the notion of TAI can play a central role in shaping strategies for developing technically successful and ethically acceptable AI systems.

### 3.3.2 AI-RELATED RISK AND UNCERTAINTY

In addition to keeping together different crucial dimensions of AI systems’ design, deployment, and assessment, the notion of TAI has further merit. Traditionally, the concepts of trust and trustworthiness have been associated with situations of risk in which the trustor is vulnerable (Nickel & Vaesen, 2012) – e.g., there is the possibility that the trustee fails to perform the delegated task. As seen in sect. “Trustworthiness and ethical requirements”, the focus on risk is pivotal in the context of AI4REALNET as well.

The notion of risk is a multifaceted one with no universally agreed-upon definition. On the one hand, non-technical understandings of risk coexist with technical ones. On the other hand, different

definitions of risk have been provided in the scientific literature. The one provided by the Royal Society in 1983, which is often referred to as the classic one, focuses on the probabilistic component of risk, which is characterized as “the probability that a particular adverse event occurs during a stated period of time or results from a particular challenge” (Royal Society, 1983).

Nowadays, definitions of risk typically involve some kind of expectation and are usually spelled out in terms of expected utility. More precisely, risk is usually defined as the combination of the probability of an unwanted event and the magnitude of its consequences (Hansson, 2009). This understanding of risk is also at the basis of the risk-based approach adopted in the AI Act, which explicitly defines risk as “the combination of the probability of an occurrence of harm and the severity of that harm” (Art. 3, 2). However, in the context of AI, and in particular in the AI Act, the notion of risk is not further articulated, and this also makes it difficult to evaluate how risk can be assessed and possibly mitigated.

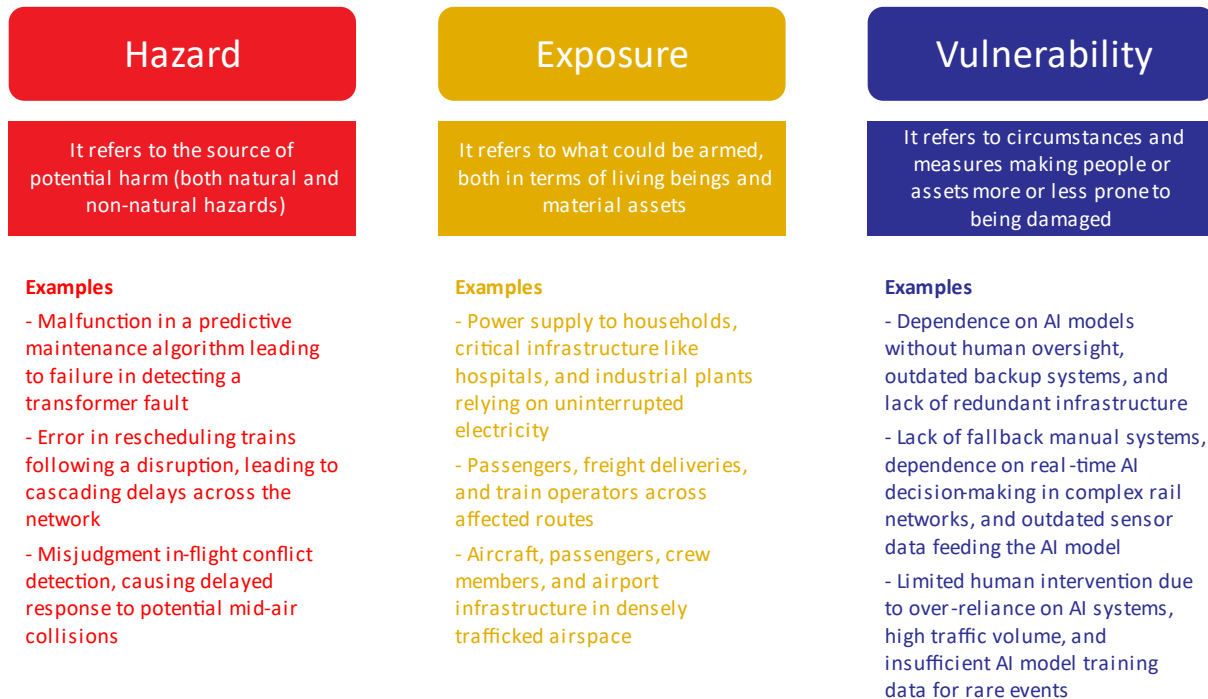
### 3.3.2.1 THE COMPONENTS OF RISK

Providing this further articulation paves the way for the specification of the methodological bases of risk assessment. In the domain of disaster risk mitigation – in particular with reference to *natural* risk management – a multi-component approach is typically adopted, decomposing risks into their different components (UNISDR, 2015):

- **Hazard** refers to the source of potential harm (e.g., a malfunction). Assessing hazard typically involves providing probabilistic estimates concerning the occurrence of the unwanted event as well as a specification of its magnitude.
- **Exposure** refers to what could be harmed as a result of the occurrence of the unwanted event. Note that exposure can concern both people and material assets, such as buildings and infrastructures.
- **Vulnerability** refers to those circumstances and features that make people and material assets more or less susceptible to the impacts of the hazard.

Risk is given by the combination of these components, which need to be all present. For instance, risks characterized by relatively low levels of hazard should be regarded as significant if many people or material assets are exposed and/or highly vulnerable. Vice versa, high levels of hazard do not automatically translate into high risks, for the components of exposure and vulnerability might be marginal.

While multi-component analyses of risk are typical in *natural* risk management, they can be fruitfully applied to the context of technological risk as well, specifically in the case of AI-related risk. It is sufficient to think about the different reasons why different AI systems might strike us as risky (Zanotti et al., 2024). Figure 40 depicts examples of the industrial domains we are considering in the AI4REALNET project.



**FIGURE 40 – AI-RELATED RISK AND ITS COMPONENTS**

This way of understanding risk provides significant advantages. First of all, it allows us to better isolate and understand the different sources of risk for a certain system. As we have seen, a system with low hazard levels, for instance, might not intuitively strike us as significantly risky. However, it might nonetheless qualify as a high-risk one due to its high levels of exposure. In addition, a multi-component focus leads us to design targeted interventions and mitigation strategies to prevent and/or mitigate the risks involved in using the system. For instance, we might want to intervene on exposure, limiting access to a certain AI-based service, or we may decide to act to reduce users' vulnerability. Importantly, this can be done on multiple fronts, simultaneously intervening in the hazard, the exposure, and the vulnerability.

### 3.3.2.2 UNCERTAINTY

Risk is often understood as being characterized by a distinctive probabilistic component: talk of risk typically implies that we can associate the potential outcomes, say, of the use of a certain technology with precise probabilities. However, this scenario is often unrealistic. It is, therefore, crucial to combine AI-related risk assessment with a rigorous analysis of the involved uncertainties (Nordstrom, 2022).

It is often hard to provide point-like probabilistic estimates in real-world risk scenarios, and AI-related risk seems to make no exception. Still, in some cases, tools such as second-order probabilities and probabilistic intervals can be used to quantify the involved uncertainties.

Things are more complicated in those contexts in which we lack solid grounds for assigning probabilities, even uncertain ones. This is especially true in the large-scale deployment of innovative technologies, for which probabilistic estimates can hardly be informed by historical data (Van de Poel, 2016). In these situations, even if we might be able to anticipate the range of potential unwanted

outcomes resulting from the introduction of a certain technology, we should acknowledge that their probabilities are characterized by hardly quantifiable forms of uncertainty.

Accordingly, the tool provided in the next pages is not meant to be employed *solely* during the design phase, for this would leave room for high degrees of uncertainty concerning the real-world use of the systems in question. Rather, it should constantly inform *regular assessments* of the systems in their final context of deployment.

### 3.3.3 A NON-CALCULATIVE TOOL FOR RISK ASSESSMENT IN SAFETY-CRITICAL SYSTEMS

Within AI4REALNET, the multi-component analysis of risk provides the conceptual and methodological grounds for applying the ALTAI to the specific context of the project. ALTAI is organized around the 7 key requirements that are at the core of the TAI framework:

1. Human Agency and Oversight;
2. Technical Robustness and Safety;
3. Privacy and Data Governance;
4. Transparency;
5. Diversity, Non-discrimination and Fairness;
6. Societal and Environmental Well-being;
7. Accountability.

Within AI4REALNET and for the safety critical systems addressed in this context, four of these requirements are particularly relevant: Human Agency and Oversight; Technical Robustness and Safety; Societal and Environmental Well-being; Accountability. The ALTAI is a useful tool for self-assessment and, despite its possible limitations, has the merit of providing a comprehensive view of the technical and ethical aspects contributing to trustworthiness. The choice of narrowing down the focus to the above-mentioned requirements is due to the fact that requirements 2, 6, and 7 seem particularly relevant in the context of safety-critical systems and can profitably be reconsidered under the lens of the multi-component analysis of risk. This analysis should also be considered when evaluating the requirement concerning Human Agency and Oversight, especially concerning the risk of overreliance on the system. This requirement is crucial for AI4REALNET due to the nature of the developed technologies, which support human decisions that should not be outsourced to the AI system.

As an example, consider the following question from the ALTAI (Requirement #2, General Safety):

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

This question, which is indeed relevant for safety critical systems, shall be further specified, decomposed, and translated into the following sub-requirements:

- **Hazard:** identify possible threats by considering both their probability of occurrence and their magnitude/impact on the system.

- **Exposure:** identify the systems’ levels of exposure to such threats, both in terms of quantity and duration.
- **Vulnerability:** implement sufficient measures to make the system less vulnerable to such threats.

Another example concerns the requirement of Societal and environmental well-being, and in particular the impact on work and skills. Even in this case, the de-skilling risk is addressed by distinguishing among (i) the affected skill and the severity of the de-skilling, (ii) the affected workforce, and (iii) the vulnerability of human operators with respect to de-skilling.

Other sub-requirements, such as the AI system’s impact on democracy (“Did you take measures that ensure that the AI system does not negatively impact democracy?”), were not considered. While different AI applications could have significant repercussions on democracy, the systems deployed within the context of AI4REALNET do not seem to constitute a direct threat in this respect.

Note that, in addition to the requirements directly stemming from the ALTAI’s questions, a requirement on risk acceptability has been added. Risk acceptability is a complex matter, and the factors making a given technological risk acceptable are highly dependent on the context and the available alternatives. It is, therefore, difficult to provide readily applicable and universally valid criteria for risk acceptability, which needs to be evaluated on a case-specific basis. As a general rule, however, it has been specified that evaluations should be made concerning the existence of alternative systems involving lower levels of risk in view of comparable positive outcomes.

Finally, an important aspect of the proposed tool is that the ALTAI questions have been reconceived in the form of positive requirements to be employed already during the design phase. The ALTAI, as a matter of fact, is mostly meant to be the basis for *ex-post* self-assessment. The tool provided here, instead, should be *proactively* employed *ex-ante*, so as to encourage active responsibility.

The key requirements, derived from the ALTAI framework and adapted for AI4REALNET’s safety-critical systems, are summarized in Table 14.

Relevant ALTAI requirement	Relevant ALTAI sub-requirement	AI4REALNET requirements
#1 Human agency and oversight	Human agency and autonomy	<ul style="list-style-type: none"> <li>• Make sure that users are adequately informed about (i) the fact that they are interacting with an AI system and (ii) the kind of inferential mechanism behind the system’s output</li> <li>• Establish mechanisms for (i) preventing over-reliance on the system and (ii) monitoring the actual use of the system to constantly check for over-reliance dynamics, especially in those scenarios for which we lack data</li> <li>• Assess the risks stemming from over-reliance by considering these risks in terms of hazard (the potential harming consequences of over-reliance), exposure (people and assets exposed to such harm), and vulnerability</li> <li>• Make sure that humans maintain <i>meaningful</i> control over the system and that their autonomy is not limited by a loss of competence due to their regularly outsourcing decisions – e.g., by blindly following recommendations – to the AI system (cf. (PRUNKL, 2022))</li> </ul>

Relevant ALTAI requirement	Relevant ALTAI sub-requirement	AI4REALNET requirements
	Human oversight	<ul style="list-style-type: none"> <li>Besides giving human operators specific training on how to exercise oversight, make sure that they are provided information on the basic working principles of RL as well as on its risks</li> </ul>
#2 Technical robustness and safety	Resilience to attacks and security	<ul style="list-style-type: none"> <li>Assess the risks stemming from potential hazards related to technical faults, outages, attacks, as well as inappropriate and malicious use</li> <li>Identify the people and material assets exposed to the potential harms resulting from such hazards</li> <li>Implement strategies to reduce the vulnerability to such hazards of (i) the system and (ii) the exposed people and assets</li> <li>Plan regular monitoring to continuously assess the involved risks and collect information on the system's real-world deployment</li> </ul>
	General safety	<ul style="list-style-type: none"> <li>Identify possible threats by considering both their probability of occurrence and their magnitude/impact on the system</li> <li>Identify the system's levels of exposure to such threats, both in terms of quantity and duration</li> <li>Implement sufficient measures to make the system less vulnerable to such threats</li> </ul>
	Accuracy	<ul style="list-style-type: none"> <li>Identify risks stemming from low levels of accuracy of the system by identifying possible hazards, the related levels of exposure, and the vulnerability of exposed people and assets, as well as measures to reduce such vulnerability</li> </ul>
	Reliability, fall-back plans and reproducibility	<ul style="list-style-type: none"> <li>Since the deployment of AI systems is often characterized by elements of uncertainty, make sure that the introduction of the system occurs in different steps, so that it is possible to evaluate risks in progressively broader controlled contexts</li> </ul>
#6 Societal and environmental well-being	Environmental well-being	<ul style="list-style-type: none"> <li>Identify the potential environmental impact of the system by considering both the training and the deployment phases</li> </ul>
	Impact on work and skills	<ul style="list-style-type: none"> <li>Assess whether and how the systematic deployment of the system might cause human de-skilling by identifying (i) the affected skills and the magnitude of the phenomenon, (ii) the affected workforce, and (iii) the contexts and features that make humans more or less prone to de-skilling, taking measures to mitigate de-skilling risks and providing training and material to enable re- and up-skilling</li> </ul>
#7 Accountability	Risk management	<ul style="list-style-type: none"> <li>Organize risk training to assure that all the three components of risk are considered</li> <li>Put in place by design mechanism in case of applications that can adversely affect individuals in terms not only of hazard but also exposure and vulnerability</li> </ul>
Additional requirements on risk acceptability:		

Relevant ALTAI requirement	Relevant ALTAI sub-requirement	AI4REALNET requirements
		<ul style="list-style-type: none"> <li>○ Given a certain system and the involved risks, make sure that there are no alternative options (with or without the use of AI) reasonably involving lower levels of risk in view of comparable positive outcomes</li> </ul>

**TABLE 14 – SUMMARY OF THE KEY REQUIREMENTS DERIVED FROM THE ALTAI FRAMEWORK AND ADAPTED FOR AI4REALNET’S SAFETY-CRITICAL SYSTEMS**



## 4. CONCLUDING REMARKS

This deliverable concludes the UCs and provides a conceptual framework description of the project. It describes six UCs across the energy and mobility domains, with high potential for human-AI teaming. From the UC description, it was possible to see that the AI-based systems in the project should be designed to raise alerts based on their confidence levels, ensuring timely human intervention while managing alert frequency to avoid operator cognitive overload. These systems allow for human override, as seen in UC1.Railway, where supervisors can take control and adjust settings based on AI confidence levels. This aligns well with the AI Act’s human oversight and intervention provisions. The co-learning process between humans and AI enables operators to request explanations and evidence, accept or reject advisories, and log interactions, allowing both the AI to learn from human preferences and the humans to improve their expertise continuously. This collaborative approach addresses potential biases and adapts to new contexts based on human feedback.

The AI system supports 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.

For this work, two tools were used in capturing requirements:

- The AI4REALNET project adapted the IEC 62559-2 standard, which defines the structure of a use case template, including lists for actors and requirements and their interrelations. This adaptation incorporated ISO/IEC TR 24030 elements to describe AI use cases, building on ISO/IEC 20547-2, IEC 62559, and IEEE P7003 standards. This approach enabled the identification of assumptions related to the business model of AI-based decision systems and their regulatory links, the description of business processes and activities, and a detailed outline of the functions supporting these processes and their associated information flows.
- Instead of using the ALTAI assessment tool as an ex-post self-assessment of AI systems, it was employed for an ex-ante assessment of UC definitions. This proactive approach fostered discussions on potential risks and ethical issues specific to the considered UCs already in the early stage of a project.

The AI4REALNET conceptual framework covers different layers, including decision process implementation and socio-technical system design, technical aspects of AI to meet requirements derived from the socio-technical level, and a transversal focus on trustworthiness from an ethical. and philosophical perspective. This framework benefits different end-users, such as AI developers, innovation managers, network operation managers, regulatory bodies, and standardization organizations, in several ways.

First, it facilitates AI development for safety-critical infrastructures by emphasizing trustworthiness, ethics, and end-user trust through various human-AI interactions and human-centered AI approaches. Additionally, by addressing multiple UCs for the operation of critical infrastructures, the framework aims to engage the AI research community, offering a broader appeal than focusing on a single use case. It also standardizes the application of AI across different critical infrastructures, ensuring consistency, quality, and compatibility of AI solutions.

Furthermore, the framework acknowledges the unique challenges and requirements of operators of critical infrastructures, providing tailored strategies and solutions while fostering collaboration among these infrastructures. It ensures that AI applications adhere to existing regulations and ethical standards, including security and transparency, starting from the design and development phase. Designed to be as technology-neutral as possible, the framework can evolve with technological advancements and changing industry requirements while also allowing for the development of AI-based systems.

## REFERENCES

- Alexander, P. A., Schallert, D. L., & Reynolds, R. E. (2009). What Is Learning Anyway? A Topographical Perspective Considered. *Educational Psychologist*, 44(3), 176–192.
- Amdouni, E., Khouadjia, M., Meddeb, M., Marot, A., Crochepierre, L., Achour, W. (2023, April). Grid2Onto: An application ontology for knowledge capitalisation to assist power grid operators. In *International Conference On Formal Ontology in Information Systems-Ontology showcases and Demos*.
- Bainbridge, L. (1983). Ironies of Automation. *Proceedings of IFAC*, 19(6), 775–779.
- Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7, 2–11.
- Behzadan, V., & Munir, A. (2017). Whatever does not kill deep reinforcement learning, makes it stronger. *arXiv preprint arXiv:1712.09344*.
- Bittner and Spence, 2003. *Use Case Modeling*, Addison-Wesley.
- Bondi, E., Koster, R., Sheahan, H., Chadwick, M., Bachrach, Y., Cemgil, T., et al, (2022, June). Role of human-AI interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 5, pp. 5286-5294)*.
- Borg, M., Bronson, J., Christensson, L., Olsson, F., Lennartsson, O., Sonnsjö, E., et al. (2021, June). Exploring the assessment list for trustworthy ai in the context of advanced driver-assistance systems. In *2021 IEEE/ACM 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEthics) (pp. 5-12)*. IEEE.
- Borst, C., Flach, J. M., & Ellerbroek, J. (2015). Beyond Ecological Interface Design: Lessons From Concerns and Misconceptions. *IEEE Transactions on Human-Machine Systems*, 45(2), 164–175. <https://doi.org/10.1109/THMS.2014.2364984>
- Bradshaw, J.M., Hoffman, R.R., Johnson, M. & Woods, D.D. (2013). The seven deadly myths of "autonomous systems". *IEEE Intelligent Systems*, 28 (3), pp. 54-61.
- Brajovic, D., Renner, N., Goebels, V. P., Wagner, P., Fresz, B., Biller, M., et al. (2023). Model reporting for certifiable AI: A proposal from merging EU regulation into AI development. *arXiv:2307.11525*.
- Braunschweig, B., Gelin, R., & Terrier, F. (2022, February). The wall of safety for AI: approaches in the confluence. *ai program*. In *Workshop on Artificial Intelligence Safety (SAFEAI)*.
- Cahour, B., Forzy, J.-F. (2009). Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science*, 47(9), 1260–1270.
- Campos, V., Klyagina, O., Andrade, J. R., Bessa, R. J., Gouveia, C. (2024). ML-assistant for human operators using alarm data to solve and classify faults in electrical grids. *Electric Power Systems Research*, 236, 110886.

- Charpentier, B., Senanayake, R., Kochenderfer, M., Günnemann, S. (2022). Disentangling epistemic and aleatoric uncertainty in reinforcement learning. arXiv preprint arXiv:2206.01558.
- Clegg, C.E. (2000). Sociotechnical principles for system design. *Applied Ergonomics* 31, pp. 463-477.
- COBBE, K., KLIMOV, O., & HESSE, C. (2019). Quantifying generalization in reinforcement learning. *International conference on machine learning*. PMLR.
- Cockburn, A., 2001. *Writing Effective Use Cases*, Addison-Wesley.
- Cremer, J. L., Kelly, A., Bessa, R. J., Subasic, M., Papadopoulos, P. N., Young, S., Marot, A., et al. (2024). A pioneering roadmap for ML-driven algorithmic advancements in electrical networks. *IEEE ISGT Europe 2024*, Zagreb, Croatia.
- De Donato, L., Flammini, F., Marrone, S., Nardone, R., Vittorini, V. (2022). Trustworthy AI for safe autonomy of smart railways: directions and lessons learnt from other sectors. *World Congress on Railway Research 2022*.
- Dignum, V. (2019, Nov.). Humane AI ethical framework. HumanE AI Deliverable 1.3. [online] <https://www.humane-ai.eu/wp-content/uploads/2019/11/D13-HumaneAI-framework-report.pdf>
- Eisbach, S., Langer, M., & Hertel, G. (2023). Optimizing human-AI collaboration: Effects of motivation and accuracy information in AI-supported decision-making. *Computers in Human Behavior: Artificial Humans*, 1(2), 100015.
- Endsley, M. R. (1988, May). Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 national aerospace and electronics conference* (pp. 789-795). IEEE
- Endsley, M. R. (2000). Situation Models: An Avenue to the Modeling of Mental Models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 44(1), 61–64. <https://doi.org/10.1177/154193120004400117>
- Endsley, M. R. (2023a). Supporting human-AI teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, 140, 107574.
- Endsley, M. R. (2023b). Ironies of artificial intelligence. *Ergonomics*, 66(11), 1656-1668.
- ENTSO-E Digital Report: The cyber physical system for the energy transition. RDIC (WG5 Digital and Communication)/ENTSO-E/POYRY 2019 [online] [https://eepublicdownloads.entsoe.eu/clean-documents/Publications/Position%20papers%20and%20reports/digital\\_report\\_2019.pdf](https://eepublicdownloads.entsoe.eu/clean-documents/Publications/Position%20papers%20and%20reports/digital_report_2019.pdf)
- Eurocontrol (2009). A White Paper on Resilience Engineering for ATM. Available at: <https://www.eurocontrol.int/sites/default/files/2019-07/white-paper-resilience-2009.pdf>
- European Commission (EC) (2024). Ethics guidelines for trustworthy AI. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Fedele, A., Punzi, C., Tramacere, S. (2024). The ALTAI checklist as a tool to assess ethical and legal implications for a trustworthy AI development in education. *Computer Law & Security Review*, 53, 105986.

- Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25(1), 3–23. <https://doi.org/10.1016/j.ijforecast.2008.11.010>
- Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1), 35–42. <https://doi.org/10.1016/j.socec.2010.10.008>
- Futia, G. and Vetrò, A. (2020). On the Integration of Knowledge Graphs into Deep Learning Models for a More Comprehensible AI. *Information*, 11 (2), 122-132.
- Gelin, R. (2024). Confiance. ai Program Software Engineering for a Trustworthy AI. In *Producing Artificial Intelligent Systems: The Roles of Benchmarking, Standardisation and Certification* (pp. 11-29). Cham: Springer Nature Switzerland.
- Gesmann-Nuisl, D., Kunitz, S. (2022). Auditing of AI in railway technology—A European legal approach. *Digital Society*, 1(2), 17.
- GOODFELLOW, I. J., SHLENS, J., & SZEGEDY, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
- Grote, G., Weik, S., Wäfler, T. & Zölch, M. (1995). Criteria for the complementary allocation of functions in automated work systems and their use in simultaneous engineering projects. *International Journal of Industrial Ergonomis*, 16, 367-382. (Download: [https://doi.org/10.1016/0169-8141\(95\)00019-D](https://doi.org/10.1016/0169-8141(95)00019-D))
- Gunning, D., & Aha, D. (2019). DARPA’s explainable artificial intelligence program. pp. vol. 40, no 2, p. 44.
- Ha, T., & Kim, S. (2023). Improving Trust in AI with Mitigating Confirmation Bias: Effects of Explanation Type and Debiasing Strategy for Decision-Making with Explainable AI. *International Journal of Human–Computer Interaction*, 1–12. <https://doi.org/10.1080/10447318.2023.2285640>
- Hackman, J. R., Oldham, G. R. (1974). Job diagnostic survey. *Journal of Applied Psychology*.
- Hackman, J. R., Oldham, G. R. (1975). Development of the job diagnostic survey. *Journal of Applied Psychology*, 60(2), 159.
- Hackman, R. J., & Oldham, G. R. (1976). Motivation through the Design of Work: Test of a Theory. *Organizational Behavior and Human Performance*, 16, 250–279.
- HANSSON, S. O. (2009). From the casino to the jungle: Dealing with uncertainty in technological risk management. *Synthese*, 168, 423-432.
- HAYES, C. F., RĂDULESCU, R., & BARGIACCHI, E. (2022). A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36, 26.
- HERNANDEZ-LEAL, P., KARTAL, B., & TAYLOR, M. E. (2019). A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33, 750-797.
- HEUILLET, A., COUTHOUIS, F., & DÍAZ-RODRÍGUEZ, N. (2021). Explainability in deep reinforcement learning. *Knowledge-Based Systems*, vol. 214, p. 106685.

- Heymann, F., Parginos, K., Bessa, R. J., & Galus, M. (2023). Operating AI systems in the electricity sector under European's AI Act—Insights on compliance costs, profitability frontiers and extraterritorial effects. *Energy Reports*, 10, 4538-4555.
- Hoffman, R. (2017). A Taxonomy of Emergent Trusting in the Human–Machine Relationship. In P. J. Smith (Ed.), *Cognitive Systems Engineering: The Future for a Changing World*. CRC Press. <https://doi.org/10.1201/9781315572529>
- Hoffman, R., Mueller, S. T., Klein, G., & Litman, J. (2018). Measuring Trust in the XAI Context (Explainable AI Program) [Technical Report]. DARPA. <https://doi.org/10.31234/osf.io/e3kv9>
- Hollnagel, E. (2015). RAG-resilience analysis grid. Introduction to the Resilience Analysis Grid (RAG). <https://www.erikhollnagel.com/onewebmedia/RAG%20Outline%20V2.pdf>
- Hollnagel, E., Woods, D.D. & Leveson, N. (Eds.). (2006). *Resilience Engineering. Concepts and Precepts*. Aldershot: Ashgate.
- Hüllermeier, E., Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3), 457-506.
- IEC 62559-2 Ed. 1.0. Use case methodology, 2014. Part 2: Definition of use case template, actor list and requirement list, 8/1340A/CDV.
- IEC/PAS 62559, 2008. IntelliGrid methodology for developing requirements for energy systems
- ILAHİ, İ., USAMA, M., & QADIR, J. (2021). Challenges and countermeasures for adversarial attacks on deep reinforcement learning. *IEEE Transactions on Artificial Intelligence*, 90-109.
- Irpan, A. (2018). Deep reinforcement learning doesn't work yet. Obtido de <https://www.alexirpan.com/2018/02/14/rl-hard.html>
- Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *arXiv*. <http://arxiv.org/abs/2010.07487>
- Jelodari, M., Amirhosseini, M. H., & Giraldez-Hayes, A. (2023). An AI powered system to enhance self-reflection practice in coaching. *Cognitive Computation and Systems*, 4(5), 243–254. <https://doi.org/DOI: 10.1049/ccs2.12087>
- Klein, G. (2018). Macrocognitive Measures for Evaluating Cognitive Work. In E. S. Patterson & J. E. Miller (Eds.), *Macrocognition Metrics and Scenarios: Design and Evaluation for Real-World Teams* (1st ed.). CRC Press. <https://doi.org/10.1201/9781315593173>
- Klein, G., & Wright, C. (2016). Macrocognition: From Theory to Toolbox. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.00054>
- Klein, G., Ross, K. G., Moon, B. M., Klein, D. E., Hoffman, R. R., & Hollnagel, E. (2003). Macrocognition. *IEEE Computer Society*, 81–84.
- Kolb, A. Y., & Kolb, D. A. (2009). The Learning Way: Meta-cognitive Aspects of Experiential Learning. *Simulation & Gaming*, 40(3), 297–327. <https://doi.org/10.1177/1046878108325713>
- Kolb, D. A. (1984). *Experimental learning: Experience as the source of learning and development*. Prentice-Hall.

- Koopman, P., & Hoffman, R. R. (2003). Work-arounds, make-work, and kludges. *IEEE Intelligent Systems*, 18(6), 70–75.
- Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*.
- Li, J., Shi, X., & Li, J. (2020). Random curiosity-driven exploration in deep reinforcement learning. *Neurocomputing*, vol. 418, 139-147.
- Li, Y. (2017). Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274 .
- Loewenstein, G., & Elster, J. (1992). Choice over time. Russell Sage Foundation.
- Lundberg, J., Johansson, B.J.E., 2021. A framework for describing interaction between human operators and autonomous, automated, and manual control systems. *Cognition, Technology & Work* 23, 381–401.
- Manzey, D. (2012). Systemgestaltung und Automatisierung. In: Badke-Schaub, P., Hofinger, G., Lauche, K. (eds.) *Human Factors*, Springer, Berlin. S. 333–352. [https://doi.org/10.1007/978-3-642-19886-1\\_19](https://doi.org/10.1007/978-3-642-19886-1_19)
- Marot, A., Donnot, B., Chaouache, K., Kelly, A., Huang, Q., Hossain, R. R., Cremer, J. L. (2022a). Learning to run a power network with trust. *Electric Power Systems Research*, 212, 108487.
- Marot, A., Kelly, A., Naglic, M., Barbesant, V., Cremer, J., Stefanov, A., & Viebahn, J. (2022b). Perspectives on future power system control centers for energy transition. *Journal of Modern Power Systems and Clean Energy*, 10(2), 328-344.
- Marot, A., Rozier, A., & Dussartre, M. (2022). Towards an AI assistant for power grid operators. *HAI2022: Augmenting Human Intellect*. IOS Press.
- Metzinger, T. (1 de 1 de 2019). Ethics washing made in Europe. *Der Tagesspiegel*, p. 1. Obtido de <https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html>
- Miller, T. (2023). Explainable AI is dead, long live explainable AI! hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 333-342).
- Mohanty, S., Nygren, E., & Laurent, F. (2020). Flatland-rl: Multi-agent reinforcement learning on trains. arXiv preprint arXiv:2012.05893.
- Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.
- Morgeson, F. P., Delaney-Klinger, K., & Hemingway, M. A. (2005). The Importance of Job Autonomy, Cognitive Ability, and Job-Related Skill for Predicting Role Breadth and Job Performance. *Journal of Applied Psychology*, 90(2), 399–406. <https://doi.org/10.1037/0021-9010.90.2.399>
- Morgeson, F. P., Humphrey, S. E. (2006). The Work Design Questionnaire (WDQ) developing and validating a comprehensive measure for assessing job design and the nature of work. *Journal of Applied Psychology*, 91(6), 1321.
- Mussi, M., Losapio, Gianvito, Maria Metelli, A., Restelli, Bessa, R.J., et al. (2024, June). Position paper on AI for the operation of critical energy and mobility network infrastructures. AI4REALNET Deliverable D2.1.

- Nachreiner, F., Nickel, P., & Meyer, I. (2006). Human factors in process control systems: The design of human–machine interfaces. *Safety Science*, 44(1), 5-26.
- Naikar, N., Brady, A., Moy, G. & Kwok, H-W. (2023). Designing human-AI systems for complex settings: ideas from distributed, joint, and self-organising perspectives of sociotechnical systems and cognitive work analysis, *Ergonomics*,66:11, 1669-1694, DOI: 10.1080/00140139.2023.2281898
- National Academies of Sciences, Engineering, and Medicine 2021. Human-AI Teaming: State of the Art and Research Needs. Washington, DC: The National Academies Press. <https://doi.org/10.17226/26355>.
- Ngo, T., & Krämer, N. (2022). I humanize, therefore I understand? Effects of explanations and humanization of intelligent systems on perceived and objective user understanding [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/6az2h>
- Nichol, A., Pfau, V., & Hesse, C. (2018). Gotta learn fast: A new benchmark for generalization in rl. arXiv preprint arXiv:1804.03720.
- Nickel, P. J., & Vaesen, K. (2012). Risk and trust (S. Roeser, R. Hillerbrand, M. Peterson & P. Sandin (Eds.), *Handbook of Risk Theory* ed.). Springer.
- Nickel, P. J., Franssen, M., & Kroes, P. C. (2010). we make sense of the notion of trustworthy technology? *Knowledge, Technology & Policy*.
- Niehaus, S., Hartwig, M., Rosen, P. H., & Wischniewski, S. (2022). An Occupational Safety and Health Perspective on Human in Control and AI. *Frontiers in Artificial Intelligence*, 5, 868382. <https://doi.org/10.3389/frai.2022.868382>
- Nikolova, I., Van Ruysseveldt, J., De Witte, H., Syroit, J. (2014). Work-based learning: Development and validation of a scale measuring the learning potential of the workplace (LPW). *Journal of Vocational Behavior*, 84(1), 1-10.
- NORDSTRÖM, M. (2022). AI under great uncertainty: implications and decision strategies for public policy. *AI & society*, 37, 1703-1714.
- Nylin, M., Johansson Westberg, J., Lundberg, J. (2022). Reduced autonomy workspace (raw)—an interaction design approach for human-automation cooperation. *Cognition, Technology & Work*, 24(2), 261-273.
- Olteanu, A., Castillo, C., & Diaz, F. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*.
- Paliouras, G. (1993). Scalability of machine learning algorithms. Doctoral dissertation, University of Manchester.
- Palminteri, S., Lebreton, M. (2021). Context-dependent outcome encoding in human reinforcement learning. *Current Opinion in Behavioral Sciences*, 41, 144-151.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52(3), 381–410.



- Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans*, 30(3), 286–297. <http://www.ncbi.nlm.nih.gov/pubmed/11760769>
- Parker, S. K., & Grote, G. (2022). Automation, Algorithms, and Beyond: Why Work Design Matters More Than Ever in a Digital World. *Applied Psychology*, 71(4), 1171–1204. <https://doi.org/10.1111/apps.12241>
- PATHAK, D., AGRAWAL, P., & EFROS, A. A. (2017). Curiosity-driven exploration by self-supervised prediction. *International conference on machine learning*. PMLR.
- Petrolo, M., G. Z., Chiffi, D., & Schiaffonati, V. (s.d.). *Two dogmas of Trustworthy AI. Model-Based Reasoning, Abductive Cognition, Creativity: Inferences & Models in Science*(Springer).
- Pohl, R. F. (2006). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (Reprint). Psychology Press.
- Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences*, 11(1), 37–43. <https://doi.org/10.1016/j.tics.2006.11.001>
- PRUNKL, C. (2022). Human autonomy in the age of artificial intelligence. *Nature Machine Intelligence*.
- Radclyffe, C., Ribeiro, M., Wortham, R. H. (2023). The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in Artificial Intelligence*, 6, 1020592.
- Rich, C., Sidner, C. L. (1997). COLLAGEN: When agents collaborate with people. In *Proceedings of the first international conference on Autonomous Agents* (pp. 284-291).
- Royal Society (1983). *Risk assessment: a study group report*. London: Royal Society.
- Royal Society (1983). *Risk assessment: a study group report*. London: Royal Society.
- RYAN, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*.
- Ryan, R. M., Deci, E. L. (2000). Intrinsic and extrinsic motivations Classic definitions and new directions. *Contemporary educational psychology*, 25(1), 54-67.
- Sadeghian, S., & Hassenzahl, M. (2022). The "Artificial" Colleague: Evaluation of Work Satisfaction in Collaboration with Non-human Coworkers. *27th International Conference on Intelligent User Interfaces*, 27–35. <https://doi.org/10.1145/3490099.3511128>
- Schaap, G., Bosse, T., & Hendriks Vettehen, P. (2023). The ABC of algorithmic aversion: Not agent, but benefits and control determine the acceptance of automated decision-making. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-023-01649-6>
- Stefani, T., Deligiannaki, F., Berro, C., Jameel, M., Hunger, R., Bruder, C., Krüger, T. (2023, October). Applying the Assessment List for Trustworthy Artificial Intelligence on the development of AI supported Air Traffic Controller Operations. In *2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC)* (pp. 1-9). IEEE.

- Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., et al., L. (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns*, 1(4).
- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185, 1124-1131. <https://doi.org/DOI: 10.1126/science.185.4157.1124>
- Ulanov, A., Simanovsky, A., & Marwah, M. (2017). Modeling scalability of distributed machine learning. *ieee 33rd international conference on data engineering*.
- United Nations Office for Disaster Risk Reduction (UNISDR) (2015). Sendai Framework for Disaster Risk Reduction 2015-2030. <https://www.undrr.org/publication/sendai-frameworkdisaster-risk-reduction-2015-2030>.
- United Nations Office for Disaster Risk Reduction (UNISDR) (2015). Sendai Framework for Disaster Risk Reduction 2015-2030. <https://www.undrr.org/publication/sendai-frameworkdisaster-risk-reduction-2015-2030>
- Van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and engineering ethics*, 22, 667-686.
- van den Bosch, K., Schoonderwoerd, T., Blankendaal, R., & Neerincx, M. (2019). Six challenges for human-AI Co-learning. In *Adaptive Instructional Systems: First International Conference, AIS 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21* (pp. 572-589). Springer International Publishing.
- Van Harmelen, F., & Ten Teije, A. (2019). A boxology of design patterns for hybrid learning and reasoning systems. *Journal of Web Engineering*, 18, 97-123.
- van Paassen, M. M., Borst, C., Ellerbroek, J., Mulder, M., & Flach, J. M. (2018). Ecological Interface Design for Vehicle Locomotion Control. *IEEE Transactions on Human-Machine Systems*, 48(5), 541–555. <https://doi.org/10.1109/THMS.2018.2860601>
- Venzke, A., Chatzivasileiadis, S. (2021). Verification of neural network behaviour: Formal guarantees for power system applications. *IEEE Transactions on Smart Grid*, 12(1), 383-397.
- Vicente, K.J., Christoffersen, K., Pereklita, A., 1995. Supporting operator problem solving through ecological interface design. *IEEE Transactions on Systems, Man, and Cybernetics* 25, 529-545.
- Von Rueden, L., MAYER, S., & BECKH, K. (2021). Informed machine learning—a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35, 614-633.
- Vouros, G. A. (2022). Explainable deep reinforcement learning: state of the art and challenges. *ACM Computing Surveys*, vol. 55, 1-39.
- Wäfler, T., & Rack, O. (2021). Kooperation und künstliche Intelligenz. In O. Geramanis, S. Hutmacher, & L. Walser (Eds.), *Kooperation in der digitalen Arbeitswelt*. Springer Fachmedien Wiesbaden. [https://doi.org/10.1007/978-3-658-34497-9\\_5](https://doi.org/10.1007/978-3-658-34497-9_5)
- Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., et al. (2023). Sociotechnical safety evaluation of generative AI systems. *arXiv:2310.11986*.

- Westin, C., Borst, C., & Hilburn, B. (2016). Strategic Conformance: Overcoming Acceptance Issues of Decision Aiding Automation? *IEEE Transactions on Human-Machine Systems*, 46(1), 41–52. <https://doi.org/10.1109/THMS.2015.2482480>
- Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A New Look at Anchoring Effects: Basic Anchoring and Its Antecedents.
- Woods, D. D. (2018). Decomposing automation: Apparent simplicity, real complexity. In *Automation and human performance* (pp. 3-17). CRC Press.
- Zanotti, G., Chiffi, D., & Schiaffonati, V. (2024). AI-Related Risk: An Epistemological Approach. *Philosophy & Technology*.
- Zanotti, G., Petrolo, M., & Chiffi, D. (2023). Keep trusting! A plea for the notion of trustworthy AI. *AI & Society*.
- Zelik, D. J., Patterson, E. S., Woods, D. D. (2010). Measuring attributes of rigor in information analysis. In E. S. Patterson & J. Miller (Eds.), *Macro-cognition metrics and scenarios: Design and evaluation for real-world teams*. Aldershot, UK: Ashgate.
- Zhang, C., Vinyals, O., & Munos, R. (2018). A study on overfitting in deep reinforcement learning. arXiv preprint arXiv:1804.06893.
- Zhang, J. M., Harman, M., Ma, L., Liu, Y. (2020a). Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 48(1), 1-36.
- Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020b). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. <https://doi.org/10.1145/3351095.3372852>
- Zissis, G. (2019). The R3 concept: reliability, robustness, and resilience [President’s Message]. *IEEE Industry Applications Magazine*, 25(4), 5-6.
- Zuboff, S. (1988). *In the Age of the Smart Machine*. BasicBooks: NewYork.

# ANNEX 1 – USE CASE TEMPLATE

**Disclaimer:** This template is an **adaption of the IEC 62559-2 standard** that defines the structure of a use case template, template lists for actors and requirements, and their relation to each other. It was a standardized template for describing use cases defined for various purposes, such as use in standardization organizations for standards development or within development projects for system development. The AI4REALNET adaptation considers the version presented in **ISO/IEC TR 24030** to describe AI use cases, which is also based on **ISO/IEC 20547-2, IEC 62559, and IEEE P7003**.

## 1 Description of the use case

A **Use Case** captures a contract between system stakeholders about its behavior. It describes the system’s behavior under various conditions as it responds to a request from one of the stakeholders, called the primary actor. Moreover, it describes the functions of a system in a technology-neutral way.

### 1.1 Name of the use case

ID	Application Domain(s)	Name of Use Case
UC.X	<b>Options:</b> Energy (power network), mobility (railway network), mobility (air traffic management)	

### 1.2 Version management

Version Management			
Version No.	Date	Name of Author(s)	Changes
0.1	DD.MM.YYYY		

### 1.3 Scope and objectives of use case

Scope and Objectives of Use Case	
<b>Scope</b>	The scope defines the limits of the use case. Example: TSO operational planning Text
<b>Objective(s)</b>	The system’s intention; what is to be accomplished; who/what would benefit. Text
<b>Deployment model</b>	Possible deployment models of AI considered in ISO/IEC TR 24030: cloud services, cyber-physical systems, embedded systems, hybrid, on-premise systems, social networks. Text

### 1.4 Narrative of use case

Narrative of Use Case	
<b>Short description</b>	
Short text intended to summarize the main idea for the reader searching for a use case or looking for an overview. <u>150 words max</u> Text	
<b>Complete description</b>	
Provides a complete narrative of the use case from a user’s point of view, describing what occurs when, why, with what expectation, and under what conditions. This narrative should be written in plain text so	

<p>non-domain experts can understand it, and can (should) have a step-by-step description. The complete description of the Use Case can range from a few sentences to a few pages. This section often helps the domain expert think through the function's user requirements before getting into the details required by the next sections of the Use Case.</p> <p>Text</p>
<p><b>Stakeholders</b></p> <p>Stakeholders that can affect or be affected by the AI system in the scenario, e.g., organizations, customers, third parties, end-users, the community, the environment, negative influencers, bad actors, etc.</p> <p>Text</p>
<p><b>Stakeholders' assets, values</b></p> <p>Stakeholders' assets and values that are at stake with potential risk of being compromised by the AI system deployment — e.g., competitiveness, reputation, trustworthiness, fair treatment, safety, privacy, stability, etc.</p> <p>Text</p>
<p><b>System's threats and vulnerabilities</b></p> <p>Threats and vulnerabilities can compromise the assets and values mentioned above – e.g., different sources of bias, incorrect AI system use, new security threats, challenges to accountability, new privacy threats (hidden patterns), etc.</p> <p>Text</p>

### 1.5 Key performance indicators (KPI)

Descriptions of KPIs for evaluating the performance or usefulness of the AI system. Descriptions include the KPI's name, description of the KPI, and reference to mentioned use case objectives. In AI4REALNET, we mention computing technical KPIs related to human and AI performance (e.g., accuracy, reward optimization, constraints satisfaction, computational time both at training and inference, attention budget).

<b>Name</b>	<b>Description</b>	<b>Reference to the mentioned use case objectives</b>
	<p>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.</p> <p>Text</p>	<p>Here is the link to one of the objectives that are specified in the targets and the KPI.</p> <p>Text</p>

### 1.6 Features of use case

<p><b>Task(s)</b></p>	<p>The main task of the use case. A pull-down list includes the following terms: recognition, natural language processing, knowledge processing and discovery, inference, planning, prediction, optimization, interactivity, recommendation and others.</p> <p>Text</p>
<p><b>Method(s)</b></p>	<p>AI method(s)/framework(s) used in development (<b>it is optional in AI4REALNET since we may leave this open</b>).</p> <p>Text</p>
<p><b>Platform</b></p>	<p>Indicate here the digital environment: Grid2Op, Flatland, BueSky.</p> <p>Text</p>

### 1.7 Standardization opportunities and requirements

<b>Classification Information</b>
<p><b>Relation to existing standards</b></p> <p>Identify here relevant standards for the use case. A good source of information:</p> <p><a href="https://www.iso.org/committee/6794475/x/catalogue/">https://www.iso.org/committee/6794475/x/catalogue/</a></p> <p><a href="https://www.etsi.org/committee/1640-sai">https://www.etsi.org/committee/1640-sai</a></p>

Text
<b>Standardization requirements</b>
Descriptions of standardization opportunities/requirements that are derived from the use case.
Text

### 1.8 Challenges and issues

<b>General Remarks</b>
Descriptions of challenges and issues of the use case.
Text

### 1.9 Societal concerns

<b>Societal concerns</b>
<b>Description</b>
Description of societal concerns related to the use case
Text
<b>Sustainable Development Goals (SGD) to be achieved</b>
The Sustainable Development Goals (SDGs), <a href="https://sdgs.un.org/goals">https://sdgs.un.org/goals</a> , are a collection of 17 global goals set by the United Nations General Assembly. SDGs are a universal call to action to end poverty, protect the planet and ensure that all people enjoy peace and prosperity. Indicate here the SGD that are within the scope of this use case.
Text

## 2 Environment characteristics

Here the nomenclature from <https://www.geeksforgeeks.org/types-of-environments-in-ai/> was used. The goal is to describe the real environment/problem (and not what currently exists in Grid2Op, Flatland or BlueSky). The information in the table below should be concise.

<b>Characteristics</b>	
<b>Observation space</b>	<p>Fully observable or partially observable? <i>Definition: "When an agent sensor is capable to sense or access the complete state of an agent at each point in time, it is said to be a fully observable environment else it is partially observable."</i></p> <p>Discrete, continuous, or mixed?</p> <p>Data update rate, e.g., 15 min data update</p> <p>Size: small (&lt; xx dimensions), medium (&gt; xx &amp; &lt;xx dimensions), large (&gt; xx dimensions)</p> <p>Text</p>
<b>Action space</b>	<p>Discrete or continuous or mixed actions?</p> <p>Size: small (&lt; xx dimensions), medium (&gt; xx &amp; &lt;xx dimensions), large (&gt; xx dimensions)</p> <p>Time horizons, e.g., next hour, all hours of the next day</p> <p>Text</p>
<b>Type of task</b>	<p>Episodic or Sequential?</p> <p><i>Definition: "In an Episodic task environment, each of the agent's actions is divided into atomic incidents or episodes. There is no dependency between current and previous incidents. In each incident, an agent receives input from the environment and then performs the corresponding action."</i></p>

	<p><i>Definition: "In a Sequential environment, the previous decisions can affect all future decisions. The agent's next action depends on what action he has taken previously and what action he is supposed to take in the future."</i></p> <p>Text</p>
<b>Sources of uncertainty</b>	<p>Deterministic or stochastic?</p> <p>Identify sources of uncertainty, e.g., weather, unplanned outages due to assets aging</p> <p>Text</p>
<b>Environment model availability</b>	<p>Physical model/equations of the environment available?</p> <p>Text</p>
<b>Human-AI interaction</b>	<p>Full-human control (AI-assisted) or co-learning (between human and AI) or full AI-based control (autonomous)?</p> <p>Text</p>

### 3 Technical details

#### 3.1 Actors

The Actor is an entity that communicates and interacts. Actors can be humans, organizations, physical objects, software applications, systems or databases, environments (physical or digital)

<b>Actor Name</b>	<b>Actor Description</b>

#### 3.2 References of use case

References (**reports, mandates and regulatory constraints, papers, patents, press releases**) associated with the Use Case and that support interest from industry and/or regulatory bodies or provide additional information from past trials/ideas. Furthermore, identify any European legal issues that might affect the design and requirements of the function, including contracts, regulations, policies, financial considerations, engineering constraints, pollution constraints, and other environmental quality issues.

<b>References</b>						
<b>No.</b>	<b>Type</b>	<b>Reference</b>	<b>Status</b>	<b>Impact on use case</b>	<b>Originator / organisation</b>	<b>Link</b>
		report, mandates and regulatory constraints, paper, patent, press release	Public / confidential	Where does the document influence the use case?		

### 4 Step-by-step analysis of use case

Template section 4 focuses on describing scenarios of the use case with a step-step analysis (sequence description). **There should be a clear correlation between the narrative and these scenarios and steps.**

#### 4.1 Overview of scenarios

The table provides an overview of the different scenarios of the use case, like normal and alternative scenarios described in section 4.2 of the template. In general, the writer of the use case starts with the

normal sequence (success). If the precondition or post-condition does not provide the expected output (e.g., no success = failure), alternative scenarios must be defined.

**In section 4.2, we consider 4 main scenarios: training, evaluation, execution, and re-training. However, it is not mandatory to use all and additional scenarios can be added.**

Scenario conditions					
No.	Scenario name	Scenario description	Triggering event	Pre-condition	Post-condition
			Event that triggers the scenario. It can be a real event (such as, “a fault occurs in the network”), or it is also possible to define scenarios that occur “periodically”.	Describes the state of the system before the scenario starts.	Describes the expected state of the system after the scenario is realized.
1	Text	Text	Text	Text	Text



## 4.2 Steps of the operational scenario 1

For this scenario, all the steps performed shall be described going from start to end using simple verbs like – get, put, cancel, subscribe, etc. Steps shall be numbered sequentially – 1, 2, 3, and so on. If needed, further steps can be added to the table (the number of steps is not limited).

<b>Scenario name</b>		<b>Training...</b>					
<b>Step No.</b>	<b>Event</b>	<b>Name of process/activity</b>	<b>Description of process/activity Service</b>	<b>Information producer (actor)</b>	<b>Information receiver (actor)</b>	<b>Information Exchanged (IDs)</b>	<b>Requirement</b>
	Event that triggers the activity.	Label for the step. Action verbs should be used when naming activity. EXAMPLE: "Fault occurs in the grid".	This describes what action takes place in this step. The focus should be less on the algorithms of the applications, and more on the interactions and information flows between actors.	Name of the actor that produces the information.	Name of the actor that receives the information.	Use an ID referring to the table in Section 5. Several IDs can be listed, comma separated.	Use an ID referring to the table in Section 6. Several IDs can be listed, comma separated.

## 4.3 Steps of the operational scenario 2

For this scenario, all the steps performed shall be described going from start to end using simple verbs like – get, put, cancel, subscribe, etc. Steps shall be numbered sequentially – 1, 2, 3, and so on. If needed, further steps can be added to the table (the number of steps is not limited).

<b>Scenario name</b>		<b>Evaluation...</b>					
<b>Step No.</b>	<b>Event</b>	<b>Name of process/activity</b>	<b>Description of process/activity Service</b>	<b>Information producer (actor)</b>	<b>Information receiver (actor)</b>	<b>Information Exchanged (IDs)</b>	<b>Requirement</b>
	Event that triggers the activity.	Label for the step. Action verbs should be used when naming activity. EXAMPLE: "Fault	This describes what action takes place in this step. The focus should be less on the algorithms of the applications, and more on the interactions and information flows between actors.	Name of the actor that produces the information.	Name of the actor that receives the information.	Use an ID referring to the table in Section 5. Several IDs can be listed, comma separated.	Use an ID referring to the table in Section 6. Several IDs can be listed, comma separated.



		occurs in the grid".					
--	--	----------------------------	--	--	--	--	--

## 5 Information exchanged

These information objects correspond to the “Information Exchanged” column referenced in the scenario steps in Section 4 “Step by Step Analysis”.

<i>Information exchanged</i>		
<i>Information exchanged (ID)</i>	<i>Name of information</i>	<i>Description of information exchanged</i>
Refers to an identifier used in the field “Information Exchanged” of Section 4.	It is a unique ID that identifies the selected information in the use case context.	Brief description, in case a reference to existing data models/information classes should be added. Using existing canonical data models is recommended.
Text	Text	Text

## 6 Requirements

This table summarizes the non-functional requirements of all steps in the Use Case and it is linked to template section 4 “Step by Step Analysis”. The ID for requirements (R-ID) is a unique ID. The following **categories** of non-functional requirements (inspired by Zhang, J. M., et al. (2020). Machine learning testing: Survey, landscapes and horizons. IEEE Trans. on Soft. Eng., 48(1), 1-36.) should be considered (but it is possible to add more):

- Robustness
- Efficiency
- Interpretability
- Regulatory and legal

<b>Requirements</b>		
<b>Categories ID</b>	<b>Category name for requirements</b>	<b>Category description</b>
Unique identifier for the category.	Name for the category of requirements.	Description of the requirement category.
<b>Requirement R-ID</b>	<b>Requirement name</b>	<b>Requirement description</b>
Unique identifier which identifies the requirement within its category.	A name of the requirement.	Description of the requirement.

## 7 Common Terms and Definitions

Follow the AI terminology and taxonomy that is currently being harmonized between EU and U.S. <https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence>

Common Terms and Definitions	
Term	Definition

# ANNEX 2 – USE CASES DESCRIPTIONS

## UC1.POWER GRID: AI ASSISTANT SUPPORTING HUMAN OPERATORS’ DECISION-MAKING IN MANAGING POWER GRID CONGESTION

### 1 Description of the use case

#### 1.1 Name of the use case

ID	Application Domain(s)	Name of Use Case
UC1.Power Grid	Power grid	AI assistant supporting human operators’ decision-making in managing power grid congestion

#### 1.2 Version management

Version Management			
Version No.	Date	Name of Author(s)	Changes
0.1	29.01.2024	Bruno Lemetayer (RTE)	Initial document (copy from last version of short template document)
0.2	01.03.2024	Bruno Lemetayer (RTE)	Process of all workshop’s feedback
0.3	05.04.2024	Bruno Lemetayer (RTE)	Preparation of final version
0.4	11.04.2024	Bruno Lemetayer (RTE)	Finalization of the document
0.5	20.04.2024	Ricardo Bessa (INESC TEC)	Non-functional requirements from ALTAI
0.6	24.04.2024	Cyrrill Ziegler (FHNW)	Insertion of Human Factors KPI’s
1.0	06.07.2024	Ricardo Bessa (INESC TEC)	Final version

#### 1.3 Scope and objectives of use case

Scope and Objectives of Use Case	
<b>Scope</b>	Power grid real-time operation and operational planning (hours-ahead)
<b>Objective(s)</b>	<p>The goal of a Transmission System Operator (TSO), and thus human operators in the control room, is to control electricity transmission on the electrical infrastructure (transmission grid) while pursuing multiple objectives, firstly to keep the system state within acceptable limits and:</p> <ul style="list-style-type: none"> <li>• keeping people and grid components safe,</li> <li>• meeting the production/consumption balance and avoid blackouts,</li> <li>• minimizing operational costs (control actions, energy losses, etc.),</li> <li>• facilitate energy transition (e.g., integration of renewables) by coping with greater uncertainty in forecasts and greater complexity of events and context.</li> </ul>

	<p>In this context, this use case describes an AI assistant that provides a human operator with recommendations for actions and/or strategies, considering the following objectives:</p> <p><u>Functional aspects</u></p> <ol style="list-style-type: none"> <li>1. Aimed at safely managing overloads on the electrical lines and, more specifically, remedial action recommendations</li> <li>2. Making the most of the renewable energies installed by limiting the emergency redispatching call to thermal power plants emitting greenhouse gases</li> </ol> <p><u>Behavioral and social aspects</u></p> <ol style="list-style-type: none"> <li>3. Easing the workload of the human operator needed to fulfill his/her missions,</li> <li>4. Integrate explainability, transparency, and trust considerations for the human operator.</li> </ol> <p>The AI assistant shall also act in a “bidirectional” manner, i.e. capitalize on the actions and the feedback from the operator with an “online” learning process running continuously.</p>
<b>Deployment model</b>	Cloud services, on-premises systems.

#### 1.4 Narrative of use case

<i>Narrative of Use Case</i>
<p><b>Short description</b></p> <p>The AI assistant oversees the transmission grid, using SCADA data and available EMS tools to identify issues and categorize them for human intervention. It monitors power flow, voltage, and balance, adhering to defined operational conditions. Anticipating problems, it sends binary alerts to the operator with confidence levels, avoiding excessive alerts to maintain operator focus (i.e., controls attention budget). Action recommendations include topological changes, storage adjustments, redispatching, and renewable energy curtailment. The human operator selects an action or seeks more information, exploring alternatives. After the operator's decision, the AI-assistant provides feedback through load flow calculations, logging decisions for continuous learning and interaction improvement.</p> <p><b>This use case only addresses congestion issues, even if other types of issues can arise on the Transmission Grid and are handled by the operators (e.g., voltage).</b></p> <p><i>Note: Different modes of interaction are possible between AI assistant and human operator, ranging from “full human control” to “full AI control”. The selected mode depends on the industry domain and context. In this use case, an ex-ante choice is made to apply a hybrid interaction where the human operator gets the final word on AI assistant recommendations.</i></p>
<p><b>Complete description</b></p> <ol style="list-style-type: none"> <li>1. The AI assistant monitors the situation of the transmission grid by using the available data from SCADA (Supervisory Control And Data Acquisition) and Energy Management System (EMS) tools and categorizes issues by distinguishing the ones needing intervention by the human operator.</li> </ol> <p>The situation of the transmission grid is monitored at the appropriate horizon (e.g., a few hours ahead to 30 minutes ahead) by using relevant forecasts (generation, consumption). Issues correspond to deviations from acceptable operation conditions of the electric system, mainly defined by:</p> <ul style="list-style-type: none"> <li>• Power flow on electric lines not exceeding thermal limits (considering, for instance, a tolerance for temporary overload).</li> <li>• Voltage maintained within a defined range.</li> <li>• Generation and load are always balanced (frequency is maintained around 50 Hz).</li> </ul>

<p>The AI assistant monitors these operating conditions and considers a predefined list of contingencies according to the operational policies of the TSO, which include:</p> <ul style="list-style-type: none"> <li>• The nominal grid, i.e., the “N” situation (in which all grid elements are available).</li> <li>• Cases in N situations where overload duration exceeds allowed thresholds: depending on TSO’s operational policies, it can be indeed allowed to let transit flows exceed a temporary threshold on a given line (e.g., flows can be higher than <math>x A</math> for 20 minutes, after which line will automatically trip). <i>Note: such equipment is used on all lines of RTE’s grid</i></li> <li>• A list of possible “N-1” (electric system’s state after the loss of one grid element and possibly several grid elements depending on the TSO’s policy).</li> </ul> <p>2. When anticipating issues requiring intervention, the AI assistant raises alerts for decisions at the appropriate horizon (e.g., a few hours ahead down to 30 minutes ahead) to the human operator in time to carry out corresponding actions. These alerts are “binary” in the sense that either the AI assistant sends a persistent alert or not, and they are associated with a level of confidence, i.e., the level of certainty of the AI assistant that the electric system won’t remain within acceptable operation conditions if no action is performed. The level of confidence is based on the uncertainty in the forecasts. The AI assistant should not send too many alerts to keep the human operator concentrated on his or her tasks and thus ease his or her workload.</p> <p>3. For a given alert, the human operator receives action recommendations from the AI assistant, with information on the predicted effect and reasons for the decision. Possible actions are:</p> <ul style="list-style-type: none"> <li>• Topological action: topology can be changed by switching power lines on and off or reconfiguring the busbar connection within substations.</li> <li>• Redispatching action: change the flexibility’s (generator, load, battery, etc.) active setpoint value. Redispatching actions include therefore storage actions (e.g., define the setpoint for charging and discharging storage units such as batteries)</li> <li>• Renewable energy curtailment: limits the power output of a given generation unit to a threshold, defined, for example, as the ratio of maximal production <math>P_{max}</math> (a value of 0.5 limits the production of this generator to 50% of its <math>P_{max}</math>).</li> </ul> <p>4. The human operator chooses a proposed recommendation or requests new information or explanations, or looks for a different action guided by an exploration agent or via manual simulation using other specific tools (that aren’t part of the AI assistant).</p> <p>5. The human operator performs needed actions according to his/her decision. The AI assistant provides feedback to the human operators on the corresponding effects: this is performed afterward (1 hour or more after the facts) by running a load flow calculation. The decisions made are logged with their corresponding context to continuously learn from realized actions and improve the interactions between the human operator and the AI assistant (e.g., relevance of proposed recommendations for actions).</p>
<p><b>Stakeholders</b></p> <p><b>TSO:</b> The transmission system operator is in charge of maintaining and operating the electricity transmission grid, which is monitored by the human operator and the AI assistant. <i>Note: This stakeholder includes all the people working for it. For example, the human operator in charge of the operation liaises with other colleagues working, e.g., in maintenance teams on the field.</i></p> <p><b>Other TSOs:</b> Neighboring TSOs are connected to the TSO via its transmission grid.</p> <p><b>Regional Control centers:</b> Control centers in charge of European operational services and TSO coordination for grid security analysis processes (e.g., TSCnet, Coreso).</p> <p><b>Human operator:</b> A member of TSO’s team who monitors the grid and takes action.</p>

**Transmission grid users:** Any party connected to the transmission grid in a contractual relationship with the TSO. This also includes Distribution System Operators (DSOs) and other critical infrastructures like railways, airports, and water treatment and distribution.

**Market participants:** Any party involved in a market whose physical underlying is electricity delivered to or from the electricity transmission grid, such as (but not limited to) wholesale markets and balancing markets.

**Stakeholders' assets, values**

**TSO, Other TSOs, Regional Control Centers**

- Legal and regulatory framework of action (e.g. Energy law defining role and missions of the TSO, European network codes).
- The AI system must enhance rather than hinder the TSO's operational competence. Risks involve misinterpretation of data, leading to incorrect decisions that impact the overall efficiency and reliability of the power transmission.
- Use of an AI Assistant by human operators must not lead to a progressive deskilling of human operators, who could lose (or won't acquire in the case of junior operators) the knowledge needed to handle more complex situations where the AI assistant can't provide any recommendation (i.e. ability to provide feedback to the AI)
- Stakeholders (in particular grid users) must trust the AI system's capabilities. Any malfunction or lack of transparency in the AI decision-making process (e.g., excessive curtailment of a renewable energy producer) can erode trust in the TSO and its ability to manage the transmission grid effectively.

It is, therefore, important to have a recurrent ex-post analysis process within TSOs to analyze the outputs of an AI system to improve confidence and also detect any bias or malfunctions.

- If the AI system's deployment is not communicated effectively or if there are public concerns regarding its use, the TSO's reputation may suffer, potentially affecting public and Energy Regulator support.  
The AI system should contribute to operational efficiency and cost-effectiveness. Moreover, the AI system's recommendations should align with sustainable energy goals.

**Human operator**

- Procedures and operation policies that define:
  - Critical boundaries, i.e., events that must be avoided (blackout or electrocution).
  - Conditions to be met by the actions (or applicable constraints/limitations), e.g., a given time must be respected between actions on a given line and changes in a generation are limited by ramp-up/down constraints.
- The human operator's decision-making authority is a significant asset. The AI system should complement human expertise.
- The integration of AI may require additional training for human operators.
- The AI system should aim to alleviate the human operator's workload rather than exacerbate it.
- The integration of AI can present opportunities for professional growth.

**Transmission Grid users**

- Depend on a reliable power supply, and the AI system must contribute to maintaining grid reliability.
- Sensitive to energy costs, and the AI system's impact on grid operations should aim to optimize efficiency and minimize operational costs.
- Expect transparency in grid operations.

**Market Participants**

The AI system's decisions should not favor specific producers unfairly, ensuring a level playing field in the energy market and promoting fair competition.

**System's threats and vulnerabilities**

**Planned and unexpected outage events:** The planned maintenance of the power grid implies that some lines are switched off for some (fixed) duration to allow their maintenance in safe conditions. Even if these events are planned and thus known in advance, they a) degrade the transmission grid's



security state and b) increase the probability of damage to the grid device (e.g., the circuit breaker used to switch back on the line). Planned events can also include regular maneuvers on grid devices to check their operating status. Grid operation can be affected by events related to equipment failures on the network (e.g., unplanned line tripping) due to aging or extreme weather events or by cyber-attacks that can disconnect the grid's equipment. Both events are external to the AI system and can increase the complexity of the solutions to solve the technical problems. The AI system will be more "exposed" to operating conditions, and the human operator will demand faster and more accurate recommendations.

**Dependency on external systems**

1) *Forecasting system*: The uncertainty of forecasts over a look-ahead horizon is intrinsically part of the base decision-making problem (or "MDP" for Markov Decision Process, which defines the environments with states and states transitions) and, therefore, part of this use case. There are several sources of uncertainty, such as weather forecast errors, interpolation errors for higher temporal resolution, or elasticity of demand to market prices. Thus, the AI-assist will make decisions under forecast uncertainty (i.e., forecast errors), which can impact its performance (e.g., generate false alerts) and require expensive corrective actions with forecast updates.

2) *SCADA measurements*: Reliance on SCADA data quality and availability in terms of nodal injections and current grid topology, which introduces vulnerabilities if those sources are compromised or unavailable.

**Adversarial data attacks**: Malicious actors might attempt to manipulate the AI system by introducing misleading data or injecting false information into the recommendation process, e.g., feeding deceptive information about the state of a particular grid node, causing it to recommend inefficient solutions or worsening congestion; or, injection of a sequence of false information to flood the human with requests during peak grid operation times.

**Trust from human operators**: The operational performance of the AI assistant will not be close to 100% of problems solved, which may hinder the confidence and trust of the human operator in the AI recommendations. This will introduce a negative cognitive bias in humans.

**Progressive deviation of environment behavior**: Not only can the system conditions evolve (production type, consumption pattern, etc.), but also the operational rules, the human operators' behavior, or other applicable regulations. This can progressively alter the efficiency of the AI assistant if it is not regularly "updated". The issue can be exacerbated by the fact that such changes happen very incrementally in time.

**A mismatch between AI training and deployment**: Related to UC2. Power Grid "*Sim2Real, transfer from simulation to real-world*", where significant differences exist between the digital environment used to train the AI model and the real operating conditions. This could lead to low robustness and poor performance during execution, e.g., recommendations based on inaccurate assumptions about grid observability and controllable resources.

**1.5 Key performance indicators (KPI)**

*Note: the table below is intended to give an exhaustive list of possible KPIs. This list will be narrowed down during the course of the project, and especially during WP4 for evaluation works.*

Name	Description	Reference to the mentioned use case objectives
Operation cost	<p>It is based on the cost of operations of a power grid that includes the cost of a blackout<sup>22</sup>, the cost of energy losses on the grid<sup>23</sup>, and the cost of remedial actions<sup>24</sup>.</p> <p>In order to simplify the computation and without hindering future improvements, it is proposed to define it as a vector whose dimensions represent different units, at least:</p> <ul style="list-style-type: none"> <li>• Number of real-time topological actions (switching actions, etc.) Only unitary actions at each timestep are considered, which means that a tuple action would be counted as two separate actions</li> <li>• Number of redispatching actions (including but not limited to storage)</li> <li>• Sum of redispatched energy volumes</li> <li>• Number curtailment action</li> <li>• Sum of curtailed energy volumes</li> <li>• Electricity losses</li> </ul> <p>Further details about cost calculation might be given during the course of the project (e.g., in WP4). This score could for example be completed with more financial aspects, such as immediate or long-term costs (e.g. indirect costs due to lifetime decay of circuit breakers). <i>Note: The cost of AI system execution is not evaluated here. See requirement E-2.</i></p>	Objectives: 1
Network utilization	<p>It is based on the relative line loads of the network, indicating to what extent the network and its components are utilized.</p> <p>This can be quantified by:</p> <ul style="list-style-type: none"> <li>• For each timestamp, the highest encountered N-1 line's load N line's load</li> <li>• The average of the maximum N-1 line's load and N line's load</li> <li>• For each timestamp, the number of lines where the N-1 line's load is greater than a given threshold (e.g., 1.0)</li> <li>• For each timestamp, the number of lines where the N line's load is greater than a given threshold (e.g., 0.9)</li> <li>• For all timestamps, the energy of overloads, calculated as the power exceeding the line capacity, integrated over the concerned timestamps (in N and N-1 state)</li> </ul>	Objectives: 1

<sup>22</sup> calculated by multiplying the remaining electricity to be supplied by the market price of electricity.

<sup>23</sup> determined by multiplying the energy volume lost due to the Joule effect by the market price of electricity.

<sup>24</sup> the sum of expenses incurred by the actions using flexibilities (e.g. balancing products, curtailment or redispatching), based on the energy volume and underlying flexibility cost.

Name	Description	Reference to the mentioned use case objectives
Topological action complexity	<p>It is used to give 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.</p> <p>Metrics for quantifying the topological utilization of the grid:</p> <ul style="list-style-type: none"> <li>• The average number of split substations (gives an indication of the distance to the reference topology)</li> <li>• The average number of substations modified in one timestamp (gives an indication of the complexity of the topological actions)</li> <li>• Number of unique split substations</li> </ul>	Objective: 1
Assistant alert accuracy	<p>It is based on the number of times the AI assistant agent is right about forecasted issues (e.g., overloads) ahead of time. Moreover, a confusion matrix can be calculated to show:</p> <ul style="list-style-type: none"> <li>• True positive cases: forecast alerts were raised by the AI assistant, and the problem did occur on the transmission grid,</li> <li>• False positive cases: forecast alerts were raised by the AI assistant, but no problem occurred on the transmission grid,</li> <li>• False negative cases: no forecast alert was raised by the AI assistant, but problems occurred on the transmission grid.</li> </ul>	Objectives: 3, 4
Assistant relevance	<p>It is based on an evaluation by the human operator of the relevance of action recommendations provided by the AI assistant and measured by the number of recommendations from the AI assistant effectively used by the human operator. It ranges in [0, 100] with:</p> <ul style="list-style-type: none"> <li>• 0 meaning that no action recommendation from the AI assistant was considered useful by the human operator,</li> <li>• 100 that all action recommendations from the AI assistant were considered useful by the human operator.</li> </ul> <p>The KPI can have values different from 0 and 100 if only a part of the action recommendations from the AI assistant were used by the human operator.</p> <p>The KPI shall distinguish between the “best decision given the information available at the time” and the “best decision in hindsight.” The evaluation shall focus on the first case, i.e., it shall not be done after the facts with full knowledge of the human operator, which was not available at the time.</p>	Objectives: 4

Name	Description	Reference to the mentioned use case objectives
Action recommendation selectivity	<p>This KPI measures how recommended actions from AI assistants contrast among KPIs used for human decisions: this allows us to put recommended actions in perspective with trade-offs used in human decisions.</p> <p>For each recommended action from the AI assistant, this KPIs consists of calculating the increase of each of the following KPIs (see above) due to action implementation:</p> <ul style="list-style-type: none"> <li>• Network utilization</li> <li>• Topological action complexity</li> <li>• Operation score</li> </ul>	Objectives: 3, 4
Assistant disturbance	<p>It aims to measure if the notifications raised by the AI assistant are disturbing the activity of the human operator. For each notification, the score ranges in [0, 5] with:</p> <ul style="list-style-type: none"> <li>• 0 meaning that the notification was not considered disturbing at all by the human operator,</li> <li>• 5 meaning that the notification was considered as fully disturbing by the human operator.</li> </ul>	Objectives: 3
Workload	It is based on a workload assessment of the AI assistant by the human operators. It shall be determined according to the NASA-TLX <sup>25</sup> methodology or similar <sup>26</sup> .	Objectives: 3
Total decision time	It is based on the time needed to decide overall, thus including the respective time taken by the AI assistant and human operator. This KPI can be detailed in a way that allows distinguishing specifically the time needed by the AI assistant to provide a recommendation.	Objectives: 3, 4
Carbon intensity	<p>It is based on the overall carbon intensity of the action recommendation, calculated as follows:</p> <ul style="list-style-type: none"> <li>• The amount of energy curtailed (or decreased following redispatching action) is split according to generation type with a negative sign</li> <li>• The amount of additional energy yielded by redispatching action is split according to generation type with a positive sign</li> <li>• The netted amount of energy <math>E_i</math> (MWh) is calculated per generation type <math>i</math></li> <li>• Each amount <math>E_i</math> is multiplied by the corresponding emission factor (kgCO<sub>2</sub>/MWh) <math>F_i</math></li> <li>• The score is then calculated as:</li> </ul> $\frac{\sum_i E_i \times F_i}{\sum_i E_i}$	Objectives: 2

<sup>25</sup> <https://humansystems.arc.nasa.gov/groups/tlx/index.php>

<sup>26</sup> See more recent works about design recommendations to create algorithms with a positive human-agent interaction and foster a pleasant user-experience: <http://hdl.handle.net/1853/61232>

<i>Name</i>	<i>Description</i>	<i>Reference to the mentioned use case objectives</i>
Trust towards the AI Tool	<p>“(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 &amp; Forzy, 2009, p. 1261). The human operators' trust towards the AI tool can be measured using the Scale for XAI (Hoffman et al., 2018) or similar.</p>	Objectives: 3, 4
Human motivation	<p>“Intrinsic motivation is defined as the doing of 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 &amp; Deci, 2000, p. 54). The human operators perceived internal work motivation can be measured by using the Job Diagnostic Survey (Hackman &amp; Oldham, 1974) or similar. The questionnaire needs to be adapted to the AI context (e.g., problem detection with AI assistant).</p>	Objectives: 3, 4
Human control / autonomy over the process	<p>“Autonomy is the degree to which the job provides substantial freedom, independence, and discretion to the employee in scheduling the work and in determining the procedures to be used in carrying it out” (Hackman &amp; Oldham, 1975, p. 162). It consists of three interrelated aspects centered on freedom in decision-making, work methods and work scheduling (Morgeson &amp; Humphrey, 2006). Parker and Grote (2022) view job autonomy interchangeably with job control. The human operator's perceived autonomy over the process can be measured by using the Work Design Questionnaire (Morgeson &amp; Humphrey, 2006) or similar. The questionnaire needs to be adapted to the AI context (e.g. problem detection with AI assistance).</p>	Objectives: 3, 4
Human learning	<p>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, as well as the continuity over time (Alexander et al., 2009). The human operators perceived learning opportunities working with the AI-based system can be measured by using the task based workplace learning scale (Nikolova et al., 2014) or similar. The questionnaire needs to be adapted to the AI context.</p>	Objectives: 3, 4

<i>Name</i>	<i>Description</i>	<i>Reference to the mentioned use case objectives</i>
Decision support for the human operator	<p>Decision support tools should be aligned with the cognitive the decision-making process that people use when making judgments and decisions in the real world and ensure that the human operator retains agency (Miller, 2023). AI decision support tools should, therefore, help people to remain actively involved in the decision-making process (e.g., by helping them critique their own ideas) (Miller, 2023).</p> <p>The decision support for the human operator can be measured based on the criteria for good decision support (Miller, 2023) or similar. The instrument needs to be further developed.</p>	Objectives: 3, 4
Ability to anticipate	<p>“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).</p> <p>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 assistant can be measured by using the Rigor-Metric for Sensemaking (Zelik et al., 2010) or similar. The instrument needs to be further developed and adapted to the AI context.</p>	Objectives: 3, 4
Situation awareness	<p>“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, p. 12).</p> <p>The human operator’s situation awareness can be measured by using the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1988) or similar.</p>	Objectives: 3, 4

**1.6 Features of use case**

<i>Task(s)</i>	Planning, prediction, interactivity, and recommendation.
<i>Method(s)</i>	Reinforcement learning has been applied to this use case, but other AI approaches are possible.
<i>Platform</i>	<a href="#">Grid2Op digital environment</a> , completed by an interactive tool allowing human operators to interact with the environment and the AI assistant

**1.7 Standardization opportunities and requirements**

<i>Classification Information</i>
<i>Relation to existing standards</i>

*ISO/IEC 23894:2023, Information technology — Artificial intelligence — Guidance on risk management.* Operating the power grid is a high-stakes task, and therefore, risk management specifically related to AI is fundamental. This standard describes the principles applied to AI, risk management framework, and processes. It is intended to be used in connection (i.e., provides additional guidance for AI) with *ISO 31000:2018, Risk management – Guidelines*.

*ISO/IEC 38507:2022, Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations.* This use case aims to augment the human operator (not only skills and knowledge but also its role), not replace him, by recognizing the complementary differences between humans and AI and leveraging them for humans. This will require an analysis of governance implications on the use of AI, namely data-driven problem-solving and adaptive AI systems (i.e., retraining during the operational phase) to new operating conditions and/or human feedback, culture, and values with respect to stakeholders, markets, and regulation.

*ISO/IEC 42001:2023, Information technology – Artificial intelligence – Management system.* This standard is the world’s first AI management system standard, providing valuable guidance for this rapidly changing field of technology. It addresses the unique challenges AI poses, such as ethical considerations, transparency, and continuous learning. For organizations, it sets out a structured way to manage risks and opportunities associated with AI, balancing innovation with governance.

*IEEE 7000-2021, IEEE Standard Model Process for Addressing Ethical Concerns during System Design.* This standard defines a framework for organizations to embed ethical considerations in concept exploration and development. It promotes collaboration between key stakeholders and ensures ethical values are traceable throughout the design process, impacting the operational concept, value propositions, and risk management. It is applicable to all organizations, regardless of size or life cycle model.

**Standardization requirements**

Application ontology that leverages agent-oriented AI recommendations to aid power grid operators in solving future problems based on past observations stored in a knowledge database. The first work in this direction was initiated in the French project CAB (Cockpit and Bidirectional Assistant), reference: Amdouni, E., Khouadjia, M., Meddeb, M., Marot, A., Crochepierre, L., Achour, W. (2023, April). Grid2Onto: An application ontology for knowledge capitalization to assist power grid operators. In International Conference On Formal Ontology in Information Systems-Ontology showcases and Demos.

In other domains of the energy sector, a good example of the use of ontologies is the Smart Applications REference (SAREF) ontology, a family of standards that enables interoperability between solutions from different providers and among various activity sectors on the Internet of Things and therefore contributes to the development of the global digital market.

**1.8 Societal concerns**

<i>Societal concerns</i>	
<b>Description</b>	
<p><b>Integration of renewable energy sources (RES):</b> Enable higher integration levels of RES and decarbonization of the economy while maintaining (or improving) the reliability and resilience of the electric power system.</p> <p><b>Resilience to extreme (natural or man-made) events:</b> Climate change is increasing the fragility of the power grid, as well as impacting the power produced by RES. Also, the digitalization of energy systems brings additional cybersecurity concerns to TSOs. These extreme events and cyber threats have not traditionally been considered in reliability standards, which typically consider reasonably probable events and neglect very improbable situations. Presently, power systems might not be sufficiently resilient to high-impact-low-probability events, which are becoming more probable.</p> <p><b>Degree of system autonomy:</b> The power grid is a critical infrastructure impacting the economy, the safety of other infrastructures, and the comfort of humans. Therefore, the type of action space is relevant, particularly if AI is providing recommendations or direct action in the environment. Furthermore, the human operator's sole ability to operate the grid and associated knowledge shall not be hampered by the AI assistant and should, on the contrary, improve thanks to interaction with the AI assistant: deskilling must be avoided.</p> <p><b>Supervision:</b> External supervision and regulator conformity assessment are present.</p> <p><b>Explainability and transparency:</b> the human operator shall be able to understand the ground basis of action recommendations provided by the AI assistant.</p>	
<b>Sustainable Development Goals (SGD) to be achieved</b>	
SGD7. Affordable and clean energy / SGD13. Climate action	

## 2 Environment characteristics

<i>Characteristics</i>	
<b>Observation space</b>	<p>Partially observable.</p> <p>Mixed: discrete (e.g., for switching device states) and continuous (e.g., for transit flows)</p> <p>Data update rate: real-time (modeled with a 5 min resolution in Grid2Op digital environment)</p> <p>Size: very large (a network with around 100 nodes has more than 4,000 dimensions. For instance, RTE's grid is composed of more than 25,000 nodes and 10,000 lines.)</p>
<b>Action space</b>	<p>Mixed actions (discrete and continuous).</p> <p>Size: large (for a network with around 100 nodes, it has &gt; 65,000 different discrete actions &amp; &gt; 200 continuous actions. For instance, RTE's grid is composed of more than 25,000 nodes and 10,000 lines.)</p> <p>All scenarios happen in an intraday time horizon, meaning not more than a 24-hour forecast period.</p>
<b>Type of task</b>	<p>Human operators and AI assistants act in a sequential environment: the previous decisions can affect all future decisions. The next action of these agents depends on what action they have taken previously and what action they are supposed to take in the future. For example, a choice of short-term remedial action can make a planned future action unavailable.</p>
<b>Sources of uncertainty</b>	<p>Stochastic (load and renewable energy forecasts, unplanned outages).</p>
<b>Environment model availability</b>	<p>Yes (physical laws of electricity).</p>



<i>Human-AI interaction</i>	Full-human control (AI-assisted) for all scenarios. Co-learning (between humans and AI) is specific to scenario 3.
-----------------------------	--

### 3 Technical details

#### 3.1 Actors

<i>Actor Name</i>	<i>Actor Description</i>
AI assistant	AI agents provide assistance to human operators. It takes information from the environment to search for recommendations and aid the human operator. In the training phase, it can act on the environment to evaluate its recommendations. In the evaluation/testing phase, the actions on the environment should be performed by the human operator only.
Human operator	A member of TSO's team is in charge of monitoring the grid and taking action on the environment (see "stakeholders" paragraph).
Environment	The human operator will interact with the Digital Environment and the AI assistant through an interface. It can be a digital environment, which is a digital model of the transmission grid, which includes unplanned events that are modeled as events appearing in predefined moments (defined directly in time series). In a real-world implementation, it is the physical environment.

#### 3.2 References of use case

<i>References</i>						
<i>No.</i>	<i>Type</i>	<i>Reference</i>	<i>Status</i>	<i>Impact on use case</i>	<i>Originator / organisation</i>	<i>Link</i>
1	Research paper	"Towards an AI Assistant for Power Grid Operators" DOI: 10.3233/FAIA220191	Public	Framework and principles for designing an AI assistant with bidirectional interactions for control room operators	Antoine Marot, Alexandre Rozier, Matthieu Dussartre, Laure Crochepierre, Benjamin Donnot	In book: HHA12022: Augmenting Human Intellect <sup>27</sup>
2	AI competition	Paris Region AI Challenge for Energy Transition, Low-carbon Grid Operations, April 2023	Public	The track "Assistant" has inspired the use case	Paris Region, RTE	Paris Region <sup>28</sup>

<sup>27</sup> [https://www.researchgate.net/publication/363763107\\_Towards\\_an\\_AI\\_Assistant\\_for\\_Power\\_Grid\\_Operators](https://www.researchgate.net/publication/363763107_Towards_an_AI_Assistant_for_Power_Grid_Operators)

<sup>28</sup> <https://www.iledefrance.fr/toutes-les-actualites/entreprises-et-chercheurs-participez-au-challenge-ia-pour-la-transition-energetique>

## 4 Step-by-step analysis of use case

### 4.1 Overview of scenarios

Notes regarding scenario and environment data:

- It is specific to scenario #1 and scenario #2.
- Scenario #3 uses scenario 1 data.

Note regarding requirements: The column “requirement” for the scenarios’ steps has been left empty for the moment. That column will get more relevant in later stages of implementation/integration when moving for a field demonstration or to demonstrate a technology with higher maturity.

Scenario conditions					
No.	Scenario name	Scenario description	Triggering event	Pre-condition	Post-condition
1	Preventive action to grant N or N-1 situation security in case of unplanned outage	<p>The AI assistant raises warnings in anticipation of the human operator and provides associated action recommendations.</p> <p>The AI assistant considers the operational context, which includes planned maintenance operations on the grid, and provides action to ensure grid security if needed.</p> <p><i>Note: a sub-scenario could address the case where the AI assistant can't provide any relevant preventive action and make this clear to the human operator, see UC2.Sim2Real.</i></p>	<p>There is a chance that the system security is not ensured at the forecasted horizon in an N or N-1 situation (for a specific case that could arise) if no action is performed.</p> <p>Thus, the AI assistant proposes actions to the operator.</p>	<p>The AI assistant continuously checks that the transmission grid security is ensured at the appropriate horizons (e.g., from a few hours ahead down to 30 minutes ahead) when considering a list of contingencies defined in the operational policies of the TSO.</p> <p>The transmission grid state (and corresponding security assessment) is forecasted.</p> <p>The Grid system is in a normal situation; there is no contingency (unexpected event on the grid), and N/N-1 situations are secured.</p>	<p>The human operator chooses one of the recommendations provided by the AI assistant.</p> <p>The transmission grid goes into the state as predicted by the AI assistant, which informs the human operator about the transmission grid state following the action performed.</p>

Scenario conditions					
No.	Scenario name	Scenario description	Triggering event	Pre-condition	Post-condition
2	AI assistant learns from human operator	The AI assistant updates its list of recommendations with actions that were performed by the human operator.	Decisions of human operators are used to improve the learning of AI assistants in new contexts.	The AI assistant is acting on new episodes that were not seen during training	All new episodes are rerun with an AI assistant trained on these new episodes. The result is compared with AI assistants not trained in these new episodes.
3	(Nice to have scenario) Human operator learns from AI assistant	The AI assistant provides feedback to the human operators on his/her actions.	The AI assistant provides feedback on actions performed by the human operator with KPIs comparing the initially recommended action and the action chosen by the operator.	Run scenario 1 from the use case Power Grid Assistant	The human operator wants to replay the scenario to get detailed feedback. The AI assistant provides feedback to the human operator on his/her actions.

#### 4.2 Steps of scenario 1

*Note: For each step, an example of operational business context is given; this will be further detailed during the definition of scenario data. Here, the scenario starts when handling a planned maintenance operation on the grid at the beginning of an operator's shift.*

Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged	Requirement
1	Start	The human operator prepares his/her shift	<p><b>Example of context:</b></p> <p>At 08:00 AM, the previous operator ended his/her shift.</p> <p>The planned outage on line L0 beginning at 09.00 AM requires 2 actions:</p> <ul style="list-style-type: none"> <li>• P1: Change topology in an adjacent substation</li> <li>• P2: Coordinate and validate a transit limitation with a DSO</li> </ul>	(empty)	(empty)	(empty)	(empty)
2	Overload forecasted	The AI assistant raises an alert	<p><b>Example of context:</b></p> <p>At 08:10, the AI assistant raises an alert for a potential overload that could occur starting at 10:00 AM on line L1 (after the N-1 situation): with the current hypothesis and forecasts, the load flow performed on the 10:00 AM situation would result in an overload.</p> <p>This overload, if confirmed, needs remedial action (else operational limits would be violated)</p> <p><i>Note: The time horizon of the scenario might need to be adjusted depending on Digital Environment's possibilities.</i></p>	AI assistant	Human operator	AIAL	(left empty)

Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged	Requirement
3	Action recommendations	The human operator processes the recommendations	<p><b>Example of context:</b> The AI assistant proposes different possible remedial actions:</p> <ul style="list-style-type: none"> <li>• A.R1: load transfer from DSO (time limit 08:15 AM)</li> <li>• A.R2: change of topology in substation S1 (time limit 09:40 AM)</li> <li>• A.R3: limitation of RES generation (costly, time limit 09:50 AM)</li> </ul> <p>AI assistant indicates A.R2 seems the best option.</p> <p><i>Note: it is more interesting to have both preventive and curative remedial actions</i></p>	AI assistant	Human operator	AIR	(left empty)
4	Time limit for remedial action R1 is reached	The AI assistant raises an alert	<p><b>Example of context:</b> At 08:15 AM, the AI assistant indicates that 1.R1's time limit is reached</p>	AI assistant	Human operator	AIAL	(left empty)
5	Operator's decision	The operator decides to ignore the recommendation R1	<p><b>Example of context:</b> The operator decides to ignore A.R1 and wait</p>	Human operator	AI assistant	D	(left empty)
6	Time limit for preparing the planned outage is reached	The AI assistant raises an alert	<p><b>Example of context:</b> The 2 actions required for planned outage beginning at 09.00 AM have to be done in time</p>	AI assistant	Human operator	AIAL	(left empty)
7	Operator's action	The operator prepares planned outage	<p><b>Example of context:</b> The operator implements action P1:</p> <ul style="list-style-type: none"> <li>• Simulation of flows with changed topology</li> <li>• Action list to change the topology</li> </ul>	Human operator	Environment	HA	(left empty)

Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged	Requirement
8	Unplanned event	An unplanned outage is needed	<p><b>Example of context:</b>            At 08:45 AM, the operator receives a call from the maintenance team.            The risk of an explosion of measuring equipment requires an urgent (ASAP) and unplanned outage.            The operator stops the ongoing actions for the planned outage to deal with the urgent outage and calls the maintenance team in charge of the planned outage to indicate that he has to stop due to another urgent outage.</p> <p><i>Note: unplanned outage could concern either:</i></p> <ul style="list-style-type: none"> <li>• a busbar: the interest is that this outage could impact in turn the list of possible remedial actions, but it might not be realistic to implement it effectively,</li> <li>• or a line, which is a simpler case.</li> </ul>	Environment	AI assistant	E	(left empty)
9	Action recommendations	The human operator processes the recommendations	<p><b>Example of context:</b>            According to the current hypothesis, the outage would result in overload in the N-1 situation at 08:50 (due to the new topology following the urgent outage).            The AI assistant proposes different possible remedial actions:</p> <ul style="list-style-type: none"> <li>• B.R1: change of topology in substation S1</li> <li>• B.R2: change of topology in substation S2</li> </ul> <p>AI assistant indicates B.R2 would make A.R2 remedial action unavailable</p>	AI assistant	Human operator	AIR	(left empty)

Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged	Requirement
10	Operator's decision	The operator decides to implement an action	<p><b>Example of context:</b> The operator goes for action B.R1</p> <p><i>Note: other combinations of cross-impacts could be imagined, for example, cases where the only possibility is that A.R2 remedial action becomes unavailable and the only possible choice is A.R1</i></p>	Human operator	AI assistant	D	(left empty)
11	Operator's action	The operator prepares unplanned outage	<p><b>Example of context:</b> The operator performs the urgent outage and implements remedial action B.R1 The operator calls the maintenance team in charge of the unplanned outage so that the urgent work can start.</p> <p><i>Note: to be detailed according to what type of grid element is concerned by the outage</i></p>	Human operator	Environment	HA	(left empty)
12	Action recommendations	The human operator processes the recommendations	<p><b>Example of context:</b> At 09:00 AM, the AI assistant proposes to continue with the remaining actions to prepare for the planned outage of line L0</p>	AI assistant	Human operator	AIR	(left empty)
13	Operator's decision	The operator decides to implement an action	<p><b>Example of context:</b> The operator decides to continue with the remaining actions to prepare planned outage of line L0</p>	Human operator	AI assistant	D	(left empty)



Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged	Requirement
14	Operator's action	The operator prepares planned outage	<p><b>Example of context:</b>            The operator confirms with DSO that action P2 can be performed            The operator implements action P2:</p> <ul style="list-style-type: none"> <li>• Topology with the simulation of agreed load transfer from DSO</li> <li>• DSO contact information</li> </ul> <p>The operator fully disconnects line L0            At 09:20 AM, the operator confirms to the maintenance team that the maintenance work can start.</p>	Human operator	Environment	HA	(left empty)
15	Time limit for remedial action R2 is reached	The AI assistant raises an alert	<p><b>Example of context:</b>            At 09:40 AM, overload is still forecasted and A.R2's time limit is reached</p>	AI assistant	Human operator	AIAL	(left empty)
16	Operator's decision	The operator decides to implement an action	<p><b>Example of context:</b>            Given that A.R2 is the only available action, the operator decides to perform A.R2</p>	Human operator	AI assistant	D	(left empty)
17	Operator's action	The operator implements an action	<p><b>Example of context:</b>            The operator implements A.R2</p>	Human operator	Environment	HA	(left empty)

#### 4.3 Steps of scenario 2

<b>Step no.</b>	<b>Event</b>	<b>Name of process/ activity</b>	<b>Description of process/ activity Service</b>	<b>Information producer (actor)</b>	<b>Information receiver (actor)</b>	<b>Information Exchanged</b>	<b>Requirement</b>
1	Start	Run episodes where the AI assistant provides recommendations	The AI assistant is acting on new episodes that were not seen during training	(empty)	(empty)	(empty)	(empty)
2	Action recommendations	The human operator processes the recommendations	(per episode) The AI assistant proposes action recommendations to the operator	AI assistant	Human operator	AIR	(left empty)
3	Operator's decision	The operator decides to implement an action	(per episode) The operator decides to take remedial action.	Human operator	AI assistant	D	(left empty)
4	Operator's preference learning	The AI assistant logs human operator's preferences	(per episode) All operator's decisions are logged for the AI assistant's learning	Human operator	AI assistant	D	(left empty)
5	Evaluation	The AI assistant's learning is evaluated	All new episodes are rerun with an AI assistant trained on these new episodes. The result is compared with the AI assistant not trained in these new episodes.	(empty)	(empty)	(empty)	(empty)

#### 4.4 Steps of scenario 3

<b>Step no.</b>	<b>Event</b>	<b>Name of process/ activity</b>	<b>Description of process/ activity Service</b>	<b>Information producer (actor)</b>	<b>Information receiver (actor)</b>	<b>Information Exchanged</b>	<b>Requirement</b>
1	Start	Run a scenario where the AI assistant provides recommendations	Use scenario 1 from the use case Power Grid Assistant	(empty)	(empty)	(empty)	(empty)

<b>Step no.</b>	<b>Event</b>	<b>Name of process/ activity</b>	<b>Description of process/ activity Service</b>	<b>Information producer (actor)</b>	<b>Information receiver (actor)</b>	<b>Information Exchanged</b>	<b>Requirement</b>
2	Operator's decision	The operator decides to implement an action	The human operator doesn't choose the remedial action recommended by the AI assistant.	Human operator	AI assistant	D	(left empty)
3	Operator's action	The operator implements an action	The operator implements the remedial action	Human operator	Environment	HA	(left empty)
4	AI assistant's instant analysis	The AI assistant provides feedback on actions performed	The AI assistant provides feedback on actions performed by the human operator with KPIs comparing the initially recommended action and the action chosen by the operator.	AI assistant	Human operator	AIAN	(left empty)
5	Replay of scenario	Go back to step #1	The human operator wants to replay the scenario	(empty)	(empty)	(empty)	(empty)
6	Action recommendations	The human operator processes the recommendations	The AI assistant provides recommendations	AI assistant	Human operator	AIR	(left empty)
7	Recommendation simulation	The human operator asks for action simulation	The human operator chooses the recommended action to see its effects / or another recommendation. The AI assistant provides simulated results of the recommended action	AI assistant	Human operator	AS	(left empty)

## 5 Information exchanged

<b>Information exchanged (ID)</b>	<b>Name of information</b>	<b>Description of information exchanged</b>
HA	Action implemented by human operator	Action (e.g., topology) implemented by human operator.
AIAL	AI assistant alert	AI assistant alerts for an overload occurring on one or several grid elements. AI assistant alert for reached time limit of a given action.
AIAN	AI assistant analysis	The AI assistant provides feedback on actions performed to the human operator.
AIR	AI assistant recommendations	List of remedial action recommended by the AI assistant
D	Decision from human operator	Human operator's choice
E	Environment information	Information on the environment, e.g., outages. <i>In case an adversarial agent is used to model unplanned events, this information would be replaced by an "adversarial attack".</i>
NRA	New remedial action	Remedial action that is not known by the AI assistant

## 6 Requirements

<b>Requirements</b>		
<b>Categories ID</b>	<b>Category name for requirements</b>	<b>Category description</b>
Ro	Robustness	It encompasses both its technical robustness (the ability of a system to maintain its level of performance under a variety of circumstances) as well as its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no unintentional harm can occur. <i>Source: EU-U.S. Terminology and Taxonomy for Artificial Intelligence. First Edition</i>
E	Efficiency	The ability of an AI system to achieve its goals or perform its tasks with optimal use of resources, including time, computational power, and data.
I	Interpretability	Make the behavior and predictions of AI systems understandable to humans, i.e., the degree to which a human can understand the cause of a decision. <i>Source: Molnar, Christoph. Interpretable machine learning. Lulu.com, 2020.</i>
Re	Regulatory and legal	The AI system's capacity to meet its objectives while complying with relevant laws, regulations, and ethical standards.
HAO	Human Agency and Oversight	The design phase involves including mechanisms for human intervention and ensuring that people can easily understand and monitor AI systems. During deployment, it means continuous monitoring and evaluation to ensure that the systems act within their ethical boundaries.

DG	Data governance	Rules, processes, and responsibilities to drive maximum value from data-centric products by ensuring applicable, streamlined, and ethical AI practices that mitigate risk and protect privacy.
FAIR	Non-discrimination and fairness	This means that AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality, and cultural diversity while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law. <i>Source: EU AI Act</i>
Acc	Accountability	Relates to an allocated responsibility. The responsibility can be based on regulation or agreement or through assignment as part of delegation. In a systems context, accountability refers to systems and/or actions that can be traced uniquely to a given entity. In a governance context, accountability refers to the obligation of an individual or organization to account for its activities, to complete a deliverable or task, to accept the responsibility for those activities, deliverables, or tasks, and to disclose the results transparently. <i>Source: EU-U.S. Terminology and Taxonomy for Artificial Intelligence. First Edition</i>
<b>Requirement R-ID</b>	<b>Requirement name</b>	<b>Requirement description</b>
Ro-1	Keep electrical grid security	The AI assistant monitors all the contingencies in the list and recommends valid actions that consider all relevant operational constraints to keep the electrical grid operating in a secure state. Thus, the physical constraints and operational limits of the electrical network should be passed to the AI system.
Ro-2	AI informs the human operator about its confidence in the output recommendation ( <i>self-awareness</i> )	Confidence of the recommendation is given by the AI assistant: Is the event really “well known” by the model thanks to its training? or is it out of distribution, and then few or no relevant recommendations can be given? The AI assistant shall indicate its confidence in the effectiveness of its recommendations with clear information, such as green, orange, or red indicators.
Ro-3	Fault tolerance	The AI system must maintain seamless grid operation despite potential failures or malfunctions within the AI infrastructure. This requires establishing robust, thoroughly tested, and efficient fallback mechanisms to ensure uninterrupted functionality.
Ro-4	Reproducibility and traceability of recommendations for <i>post-mortem</i> analysis	All recommendations made by the AI system must be reproducible at a later point, given the same input or specific context/conditions. While the actions recommended by the system do not need to be identical in a strict mathematical sense - acknowledging the variability inherent in distributed computing environments - they should be closely aligned and functionally equivalent, ensuring reliable and predictable outcomes under similar conditions. Moreover, it should be possible to trace back which AI model or rules led to the decision(s) or recommendation(s) of the AI system, which is very relevant for audits from the Energy Regulator.

Ro-5	Adaptability to different operating conditions	The system should be able to adapt to different scenarios or operational conditions without significant degradation in performance (i.e., maintain appropriate levels of stability). The scenarios considered are related to the training examples but are particularly challenging.
Ro-6	Do not increase cybersecurity risk	The AI assistant should not increase the system's overall cybersecurity risk level. It must be closed to adversarial attacks from external parties so that no control is taken over the information provided to the human operator. It must also be designed to prevent any communication with commands of grid components (e.g., opening of circuit breakers).
Ro-7	Keep acceptable performance levels under natural or adversarial perturbations during operation	The training of the AI system should include scenarios with natural or adversarial perturbations in its input/state vector, which can originate from missing or erroneous values from the environment ( <i>or adversarial attacks from agents</i> ).
Ro-8	Robustness to attacks targeting model space and reward function	Reward functions and models should be stored and operated in highly cyber-secure Information Technology (IT) systems. In the event of an attack, the previously trained model could be quickly restored. Model training should be done in a secure and controlled digital environment, and model retraining is possible.
Ro-9	Detect changes in AI behavior	Changes in the AI system should be auditable and controlled by humans. Nevertheless, several supervised and reinforcement learning algorithms have online learning, and it might be difficult to evaluate or detect changes in the AI system. Thus, automatic mechanisms are required to detect data and model shifts.
E-1	Relevance of the recommendations	The AI assistant often becomes confident in its ability to propose relevant recommendations to solve situations and limits its number of warnings to the human operator to help him focus his/her attention.
E-2	Computational efficiency	The AI system must be designed to ensure efficient training and inference capabilities on various computer hardware, from small-scale development setups with limited processing power to configurations involving multiple servers and GPUs.
E-3	Scalability	The AI system's training and inference methodology and algorithms must be designed to scale up for applications in large and realistic electrical networks.
E-4	Adequate training environment	AI-friendly digital environments should be used to train the AI system, which generates high-quality representative data of the environment where the system will be deployed. However, the transfer of knowledge from simulation to the real environment should be carefully designed – see UC2.Power Grid “Sim2Real, transfer AI-assistant from simulation to real-world operation”.
I-1	Action rating	Frame recommendations into different scenarios/strategies, and rate these scenarios based on their consequences, e.g., identify a “robust” strategy that could work in all cases or a “no regret” strategy.

I-2	Transparency during system training	The AI system must exhibit high transparency in its decision-making processes. This necessitates that documentation on the system's training data, training methods, and scenarios is available and understandable to relevant stakeholders.
I-3	Capacity to explain recommendation(s) to the human operator (and other stakeholders)	Depending on the type of AI model used, different options are possible, such as (non-exhaustive list): a) empirically compare the outcomes of various strategies and evaluate the proposed recommendations against predefined KPIs; b) relate the recommendations with features importance of the state/input vector; c) use inherently interpretable models and/or knowledge distillation to explain the decisions of a more complex/large model. A trade-off between accuracy and interpretability needs to be evaluated.
I-4	Adaptability to different levels of interaction and human operator preferences and experience	Each operator has its own preferences (e.g., one operator can be more risk averse than others): ideally, the AI assistant interacting with a given operator could provide decision support that fits the preferences of this operator but is not necessary of another, especially given the type of situation that can require more attention. Thus, the AI system shall be able to interact with the human operator according to his/her preferences and experience, such as a) fully manual, b) get notified every time an overload is detected, and c) only get notified when the AI assistant is not confident enough.
Re-1	Compliance with existing operational policies	The AI assistant's recommendations comply with operational policies and network codes for power grids.
Re-2	European AI Act	The AI system must be prepared to comply with the regulations and standards stipulated in the European AI Act. This compliance involves adhering to the defined transparency, safety, data governance, and accountability requirements.
Re-3	Transparency to humans in terms of interaction with an AI system	The human operator should be aware of their interaction with an AI or another human. In this case, operators are advised of the AI assistant and, hence, not be confused about whether they interact with a human or AI system.
Acc-1	Allow audits for the AI recommendations and human operator actions	Audits are to be expected, though no formal assessment process is available for software in the power grid domain. The regulator will look at the case if a grid user or electricity market agent has a complaint. This is strongly related to requirements Ro-4 and I-3.
Acc-2	Reporting of potential vulnerabilities, risks, or biases	A database with vulnerabilities, risks, and biases, similar to <a href="#">AI Vulnerability Database</a> should be created. However, the vulnerabilities and risks of other systems, e.g., SCADA, should be evaluated together due to interdependencies with the AI system (e.g., source of input data).
HAO-1	Mitigate addictive behavior from humans	The AI system should operate as a recommender (i.e., one more additional tool to support the human operator's decisions), and all the decisions should be solely taken by the human operator (human-in-command approach). The AI assistant shall not create a craving among the operators to use it. On the other hand, we should maintain credibility and intimacy between the operator and the AI system.

HAO-2	Mitigate de-skilling in the human operators	The usage of the AI system must not lead to de-skilling in the human operators. This requires new metrics that monitor workers' skill levels and provisions for actions to compensate workers' de-skilling. Furthermore, a higher knowledge of the fundamentals behind the AI system can help human operators understand the decision-support process.
DG-1	Processing of human operator data	The AI system can use historical data about human operator actions, employing techniques such as imitation learning. However, it is imperative that this data undergoes complete anonymization, as the identification of individual operators is unnecessary. Including action timestamps is mandatory, ensuring compatibility with a table of operator shifts. Consequently, even when cross-referenced, it should remain impossible to discern the operator's identity or correlate specific actions with individuals (including performance metrics). Additionally, the knowledge database must exclude any actions characterized by poor performance.
FAIR-1	Avoid creating or reinforcing unfair bias in the AI system	The system must not unfairly favor specific producers or consumers of electrical energy. A level playing field in the electricity market, as well as fair competition, must be provisioned. Measures must be implemented to ensure these fairness constraints are observed. Note that: 1) Occurring bias may very well originate from technical or physical limitations of electrical grid operations and hence may (in part or wholly) not be avoidable. 2) Requiring the AI system to adhere to fairness standards that are not required from existing alternative techniques may put it at a disadvantage, especially if those originate from the source of the previous issue.
FAIR-2	Regular monitoring of fairness	Using the physical equations of the power grid, it is possible to compare the decisions made by the AI system and the impact that other grid users would have in solving the technical problem. For instance, <i>ex-post</i> , it is possible to run an optimal power flow with the redispatch costs and compare its solution with the AI system. Having a least-cost solution is the primary goal. Metrics such as Jain's fairness index have been used to evaluate fairness in load shedding <sup>29</sup> and fairness in renewables' curtailment <sup>30</sup> .

## 7 Common Terms and Definitions

Common Terms and Definitions	
Term	Definition
TSO – Transmission System Operator	A natural or legal person is responsible for operating, ensuring the maintenance of, and, if necessary, developing the transmission system in a given area and, where applicable, its interconnections with other systems and for ensuring the long-term ability of the system to meet reasonable demands for the transmission of electricity. Source: Directive 2009/72/EC and ENTSOE glossary

<sup>29</sup> F. Moret and P. Pinson, "Energy Collectives: A Community and Fairness Based Approach to Future Electricity Markets," IEEE Trans. Power Syst., vol. 34, no. 5, pp. 3994–4004, Sep. 2019.

<sup>30</sup> M. Z. Liu Liu, A. T. Procopiou, K. Petrou, L. F. Ochoa, T. Langstaff, J. Harding, and J. Theunissen, "On the Fairness of PV Curtailment Schemes in Residential Distribution Networks," IEEE Trans. Smart Grid, vol. 11, no. 5, pp. 4502–4512, 2020.



SCADA - Supervisory Control And Data Acquisition	A system of different hardware and software elements that come together enables a power grid operator to monitor and control various components of a power system in real time, such as generators, transformers, and transmission lines.
EMS – Energy Management System	Optimal control center solution enables secure, efficient, and optimized electric power system operation.
Nominal grid (“N” situation)	Network operating condition where all grid elements are available
Contingency (“N-1” situation)	Electric system’s state after the loss of one grid element, and possibly several grid elements, depending on the TSO’s policy
Load (or power) flow calculation	Calculations are used to determine the voltage, current, and real and reactive power at various points in a power system under steady-state conditions.
Line’s load	It is defined as the observed current flow divided by the thermal limit of each powerline (no unit): the value is within [0; 1] interval. A line’s load is associated with a line for a given state: it is therefore referred to as “N line’s load” or “N-x line’s load”. <i>Note: this measure is referred to as “rho” in Grid2Op digital environment</i>

## UC2.POWER GRID: SIM2REAL, TRANSFER AI-ASSISTANT FROM SIMULATION TO REAL-WORLD OPERATION

### 1 Description of the use case

#### 1.1 Name of the use case

ID	Application Domain(s)	Name of Use Case
UC2.Power Grid	Power grid	Sim2Real, transfer AI-assistant from simulation to real-world operation

#### 1.2 Version management

Version Management			
Version No.	Date	Name of Author(s)	Changes
0.1	29.01.2024	Bruno Lemetayer (RTE)	Initial document (copy from last version of short template document)
0.2	01.03.2024	Bruno Lemetayer (RTE)	Process of all workshop's feedback
0.3	05.04.2024	Bruno Lemetayer (RTE)	Preparation of final version
0.4	11.04.2024	Bruno Lemetayer (RTE)	Finalization of the document
0.5	20.04.2024	Ricardo Bessa (INESC TEC)	Final review and inclusion of non-functional requirements
0.6	24.04.2024	Cyrill Ziegler (FHNW)	Insertion of Human Factors KPI's
1.0	08.07.2024	Ricardo Bessa (INESC TEC)	Final version

#### 1.3 Scope and objectives of use case

Scope and Objectives of Use Case	
<b>Scope</b>	Power grid real-time operation and operational planning (hours-ahead). It integrates the global concept of the assistant framework (developed in UC1.Power Grid) and deepens a specific "real world" complication (in comparison, UC1.Power Grid has a more "theoretical" vision).
<b>Objective(s)</b>	<p>This use case is to assess the capability of an AI assistant to be used for the operation of a "real" transmission grid, in the sense that the "real" environment doesn't exactly behave as the one available to the agent (that is implemented in the AI assistant) during training and simulation procedures, even if they share the same functional properties (same grid components and topology), and operational constraints. Therefore, Sim2Real stands for "Simulation to Reality".</p> <p>The main objectives are:</p> <ol style="list-style-type: none"> <li>1. Look at additional technical considerations to succeed at deploying an AI assistant in the real world besides its sole ability to find solutions to simulated situations.</li> <li>2. Improving human trust when such systems are deployed in real-world environments.</li> </ol>

	3. Allowing for iterative human-AI refinements with human feedback and insights.
<b>Deployment model</b>	Possible deployment models of AI considered in ISO/IEC TR 24030: cloud services, on-premises systems,

**1.4 Narrative of use case**

<b>Narrative of Use Case</b>	
<b>Short description</b>	
<p>The use case outlines two paths for an AI assistant to manage a transmission grid:</p> <p>A) In coping with real-world conditions, the AI assistant monitors grid situations, raises alerts for human intervention, and provides action recommendations, considering uncertainty coming from bad or low-quality data (e.g., partially missing). The human operator makes decisions based on AI suggestions, with feedback loops to continuously improve interactions and learn from realized actions.</p> <p>B) When data limitations prevent full autonomy, the AI assistant alerts the human operator due to missing or poor-quality data. The human operator may also choose actions that do not yield expected results due to various factors. In such cases, the operator can provide missing information to aid the AI. Enriched context, including human input and decisions, is logged for continuous learning, enhancing the AI assistant’s robustness in making recommendations for grid actions.</p> <p>This use case only addresses congestion issues, even if other types of issues can arise on the Transmission Grid and are handled by the operators (e.g., voltage).</p> <p><i>Note 1: This use case is linked with the broader notion of “transfer learning”, which is the possibility to adapt a pre-trained model to a new environment only with a slight additional training. One of the possible associated research questions is to evaluate the minimum amount of real data that would be needed to align a model with the “real world”. In the context of this use case, transfer learning won’t be applied, and the model trained in the context of the Power Grid Assistant use case will be used.</i></p> <p><i>Note 2: As for the AI-assistant training, the human operator’s decision and perception will rely on “theoretical simulations” (training and simulation tools).</i></p>	
<b>Complete description</b>	
<p>The use case can be divided into two paths:</p> <p><b>A. The AI assistant copes with real-world conditions</b></p> <p>The AI assistant can still carry out its role and provide the human operator with action recommendations, even if data is not of good quality as in training.</p> <ol style="list-style-type: none"> <li>6. The AI assistant monitors the transmission grid situation <i>[same as in UC1.Power Grid]</i></li> <li>7. When anticipating issues requiring intervention, the AI assistant raises alerts for decisions at the appropriate horizon (e.g., a few hours ahead to 30 minutes ahead) to the human operator in time for carrying out corresponding actions <i>[same as in UC1.Power Grid]</i> The action recommendations from the AI assistant will reflect the additional uncertainty due to bad-quality data and the sensitivity to uncertainty.</li> <li>8. For a given alert, the human operator receives action recommendations from the AI assistant, with information on the predicted effect and reasons for the decision <i>[same as in UC1.Power Grid]</i></li> <li>9. The human operator chooses a proposed recommendation, or requests new information or explanations, or looks for a different action guided by an exploration agent or via manual</li> </ol>	

simulation using other specific tools (that aren't part of the AI assistant) *[same as in UC1.Power Grid]*

10. The human operator performs needed actions according to his/her decision *[same as in UC1.Power Grid]*
11. The decisions made are logged with their corresponding context to continuously learn from realized actions and improve the interactions between the human operator and the AI assistant (e.g., relevance of proposed recommendations for actions) *[same as in UC1.Power Grid]*

### **B. Real-world conditions require specific interactions between AI assistant and human operator**

Available data doesn't allow the AI assistant to provide the human operator with action recommendations in a fully autonomous way and requires the AI assistant to call for additional feedback or information from the human operator: the AI assistant raises an inaccuracy alert.

1. **The first type of situation is where the AI assistant can't evaluate the need for action due to missing and bad-quality data** and thus can't determine any action recommendations. It raises a corresponding alert to the human operator.

The main reasons can be:

- Bad or low-quality data:
  - Due to uncertainty because the forecasts aren't always accurate or even available, or uncertainty as "epistemic uncertainty", which is the model uncertainty due to sampling (or underrepresentation) problems
  - The state estimator does not directly use the measurement values but first goes through a readjustment. This means that the raw measurement values from the Energy Management System (EMS) can't be directly used to compute the load flow because the needed adaptations (missing or wrong measurement values due to, e.g., measurement device issues) performed by the state estimator will be missing.
- Evolution of the electric system: trends such as higher renewable penetration or consumer behavior change (adaptation) that shift data distribution over the years.

2. **The second type of situation** is where a recommended action doesn't have the expected consequences on the transmission grid's state.

The main reasons can be:

- Reproducibility of remedial actions, one or several prerequisites needed to perform an action recommended by the AI-assistant are missing due to:
  - Device failure (e.g., the failure of a circuit breaker might prevent changing the topology as proposed).
  - Unavailability of flexibility (that might prevent performing planned redispatching).
- Real-time behavior of the transmission grid is significantly different from simulation due to:
  - Different load flow calculation than the one available at training and inference time.
  - Add or upgrade new elements on the grid: substations, lines, etc., even automatic devices.
  - Distributed energy resources (DER) can impact grid congestion and decision-making since they can be a source of additional complexity and difficulty: a model might not be able to analyze or predict the real-world cumulative effect of smaller grid-connected assets.
  - changing grid equipment characteristics (e.g., climate impact or DLR).

<ul style="list-style-type: none"> <li>○ transient grid dynamics that steady state simulation doesn't capture, for example, in the context of a windstorm.</li> <li>○ cyber-physical considerations with the integration and modeling of more automatic devices.</li> </ul> <p>3. When the AI assistant can't evaluate the need for action, or a recommended action doesn't have the expected consequences, the human operator can provide the AI assistant with specific missing information to help the AI assistant forecast system state and assess action recommendations. This is only possible if the human operator can easily provide missing information to the AI assistant (i.e., it doesn't generate an important additional workload), e.g., the status (open/closed) of a given busbar coupler.</p> <p>4. The difference between the original context used by the AI assistant and the enriched context is logged to continuously learn from realized actions and improve the robustness and novelty of recommendations for actions by the AI assistant. Enriched context includes at least:</p> <ul style="list-style-type: none"> <li>• information given by the human operator.</li> <li>• Decisions are made by the human operator (visible as topology changes or other actions on the transmission grid).</li> </ul>
<b>Stakeholders</b>
See UC1.Power Grid
<b>Stakeholders' assets, values</b>
See UC1.Power Grid
<b>System's threats and vulnerabilities</b>
<p><b>Human manipulation:</b> Human operators with malicious intent may attempt to manipulate the AI system by providing misleading feedback or deliberately misusing the AI learning process. It is important to ensure that this co-learning process complies with regulatory requirements and industry standards for power grid management.</p> <p><b>Adversarial data attacks:</b> Malicious actors might attempt to manipulate the AI system by introducing misleading data or injecting false information into the recommendation process, e.g., feeding deceptive information about the state of a particular grid node, causing it to recommend inefficient solutions or worsening congestion; or, injection of a sequence of false information to flood the human with requests during peak grid operation times.</p> <p><b>Trust from human operators:</b> The operational performance of the AI assistant will not be close to 100% of problems solved, which may hinder the confidence and trust of the human operator in the AI recommendations. This will introduce a negative cognitive bias in humans.</p>

### 1.5 Key performance indicators (KPI)

*Note: the table below is intended to give an exhaustive list of possible KPIs. This list will be narrowed down during the course of the project, and especially during WP4 for evaluation works.*

<i>Name</i>	<i>Description</i>	<i>Reference to the mentioned use case objectives</i>
Technical robustness to real-world imperfections	<p>Describes the ability of the AI system to maintain its performance level under natural or adversarial perturbations, namely bad or low-quality data, or when recommended action does not have the expected impact on the transmission grid's state. This KPI can be quantified by comparing the technical performance of the AI assistant without and with the perturbations, using KPIs from UC1.Power Grid. From those KPIs, the following metrics (or properties) can be computed:</p> <ol style="list-style-type: none"> <li>1) The extent to which the output of the AI system or a specific KPI (e.g., operation score) varies with the perturbations, e.g., measured with the output/KPI variance and/or average difference.</li> <li>2) Assess whether a particular decision holds for input variation (data quality issue) in the same context.</li> </ol> <p>During the training-time of the AI assistant, the slope of the reward/loss function deterioration can also be used to measure technical robustness.</p>	Objectives: 1,2,3
Resilience to real-world imperfections	<p>Ability to prepare for and adapt to changing conditions and withstand and recover (to a "normal" state) rapidly from natural or adversarial perturbations or unexpected changes. The quantification of this KPI can be made with the magnitude and/or duration of reward/loss function performance degradation compared to an unperturbed system for the same context. It can, for instance, be measured by the area between the reward curves of the unperturbed and perturbed AI system. This can be computed during training or operational testing time.</p>	Objectives: 1,2,3
Transferability across fidelity levels	<p>Measures how effectively a policy or model trained in one environment (low-fidelity simulation) performs when applied to different environments (e.g., high-fidelity simulation or real-world operation). Evaluated by directly applying the policy trained in a low-fidelity simulation to a high-fidelity simulation and measuring its effectiveness by computing the KPIs from UC1.Power Grid.</p>	Objectives: 1,2,3
Generalization to different grid operating conditions	<p>The ability of a policy to perform well in an unseen grid operation condition that was not part of the training experience. Tested by exposing the previously trained AI system to different environments with changed grid elements and observing how well it adapts and performs by determining the KPIs from UC1.Power Grid.</p>	Objectives: 1,2,3
Assistant disturbance	<p>It aims to measure if the notifications raised by the AI assistant are disturbing the activity of the human operator. For each notification, the score ranges in [0, 5] with:</p> <ul style="list-style-type: none"> <li>• 0 meaning that the notification was not considered disturbing at all by the human operator,</li> <li>• 5 meaning that the notification was considered as fully disturbing by the human operator.</li> </ul>	Objectives: 3

Name	Description	Reference to the mentioned use case objectives
Workload	It is based on a workload assessment of the AI assistant by the human operators. It shall be determined according to the NASA-TLX <sup>31</sup> methodology or similar <sup>32</sup> .	Objectives: 3
Assistant self-awareness	<p>It is based on the number of times the AI assistant agent is right about its ability to perform action recommendations ahead of time. Moreover, a confusion matrix can be calculated to show:</p> <ul style="list-style-type: none"> <li>• True positive cases: AI assistant raises inaccuracy alert indicating it has insufficient data to estimate the state of the grid and it actually doesn't have the required data,</li> <li>• False positive cases: AI assistant raises inaccuracy alert indicating it has insufficient data to estimate the state of the grid, but it actually does have the required data (i.e., it should be confident, but it isn't)</li> <li>• False negative cases: AI assistant doesn't raise inaccuracy alert, but in reality, it can't properly assess the situation (i.e., is falsely confident)</li> </ul> <p><i>Note: This KPI is the adaptation of the "Assistant alert accuracy" KPI of UC1 "Power Grid Assistant"</i></p>	Objectives: 3
Trust towards the AI Tool	<p>"(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 &amp; Forzy, 2009, p. 1261).</p> <p>The human operators' trust towards the AI tool can be measured using the Scale for XAI (Hoffman et al., 2018) or similar.</p>	Objectives: 2,3
Human motivation	<p>"Intrinsic motivation is defined as the doing of 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 &amp; Deci, 2000, p. 54).</p> <p>The human operators perceived internal work motivation can be measured by using the Job Diagnostic Survey (Hackman &amp; Oldham, 1974) or similar. The questionnaire needs to be adapted to the AI context (e.g., problem detection with AI-assistance).</p>	Objectives: 2,3

<sup>31</sup> <https://humansystems.arc.nasa.gov/groups/tlx/index.php>

<sup>32</sup> See more recent works about design recommendations to create algorithms with a positive human-agent interaction and foster a pleasant user-experience: <http://hdl.handle.net/1853/61232>

<i>Name</i>	<i>Description</i>	<i>Reference to the mentioned use case objectives</i>
Human control/ autonomy over the process	<p>“Autonomy is the degree to which the job provides substantial freedom, independence, and discretion to the employee in scheduling the work and in determining the procedures to be used in carrying it out” (Hackman &amp; Oldham, 1975, p. 162). It consists of three interrelated aspects centered on freedom in decision-making, work methods and work scheduling (Morgeson &amp; Humphrey, 2006). Parker and Grote (2022) view job autonomy interchangeably with job control.</p> <p>The human operator's perceived autonomy over the process can be measured by using the Work Design Questionnaire (Morgeson &amp; Humphrey, 2006) or similar. The questionnaire needs to be adapted to the AI context (e.g., problem detection with AI assistance).</p>	Objectives: 2,3
Human learning	<p>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, as well as the continuity over time (Alexander et al., 2009).</p> <p>The human operators perceived learning opportunities working with the AI-based system can be measured by using the task based workplace learning scale (Nikolova et al., 2014) or similar. The questionnaire needs to be adapted to the AI context.</p>	Objectives: 2,3
Decision support for the human operator	<p>Decision support tools should be aligned with the cognitive the decision-making process that people use when making judgments and decisions in the real world and ensure that the human operator retains agency (Miller, 2023). AI decision support tools should therefore help people to remain actively involved in the decision-making process (e.g. by helping them critique their own ideas) (Miller, 2023).</p> <p>The decision support for the human operator can be measured based on the criteria for good decision support (Miller, 2023) or similar. The instrument needs to be further developed.</p>	Objectives: 2,3



<i>Name</i>	<i>Description</i>	<i>Reference to the mentioned use case objectives</i>
Ability to anticipate	<p>“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).</p> <p>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 an AI-based assistant can be measured by using the Rigor-Metric for Sensemaking (Zelik et al., 2010) or similar. The instrument needs to be further developed and adapted to the AI context.</p>	Objectives: 2,3
Situation awareness	<p>“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, p. 12).</p> <p>The human operator’s situation awareness can be measured by using the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1988) or similar.</p>	Objectives: 2,3

### 1.6 Features of use case

<i>Task(s)</i>	Planning, prediction, interactivity, recommendation, inference.
<i>Method(s)</i>	Reinforcement learning has been applied to this use case, but other AI approaches are possible.
<i>Platform</i>	<a href="#">Grid2Op digital environment</a> , completed by an interactive tool allowing human operators to interact with the environment and the AI assistant

### 1.7 Standardization opportunities and requirements

<i>Classification Information</i>
<b>Relation to existing standards</b>
<p><i>ISO/IEC 23894:2023, Information technology — Artificial intelligence — Guidance on risk management.</i> Operating the power grid is a high-stakes task, and therefore, risk management specifically related to AI is fundamental. This standard describes the principles applied to AI, risk management framework, and processes. It is intended to be used in connection (i.e., provides additional guidance for AI) with <i>ISO 31000:2018, Risk management – Guidelines</i>.</p> <p><i>ISO/IEC 24029-2:2023, Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for using formal methods.</i> Artificial neural networks are generally a building block of AI assistants for power grid operation (see results from L2RPN competitions); thus, methodologies for using formal methods to assess the robustness properties of neural networks are important. This standard is focused on how to select, apply, and manage formal methods to prove robustness properties. The technical report <i>ISO/IEC TR 24029-1:2021</i> complements this standard and presents an overview of different methods to assess the robustness of neural networks.</p> <p><i>ISO/IEC 42001:2023, Information technology – Artificial intelligence – Management system.</i> This standard is the world’s first AI management system standard, providing valuable guidance for this rapidly changing field of technology. It addresses the unique challenges AI poses, such as ethical considerations, transparency, and continuous learning. For organizations, it sets out a structured way to manage risks and opportunities associated with AI, balancing innovation with governance.</p> <p><i>IEEE 7000-2021, IEEE Standard Model Process for Addressing Ethical Concerns during System Design.</i> This standard defines a framework for organizations to embed ethical considerations in concept exploration and development. It promotes collaboration between key stakeholders and ensures ethical values are traceable throughout the design process, impacting the operational concept, value propositions, and risk management. It is applicable to all organizations, regardless of size or life cycle model.</p>
<b>Standardization requirements</b>
<p>Assessment of AI robustness should go beyond artificial neural networks (ISO/IEC 24029-2:2023) and consider other AI models, as well as the communication of this information to the end-user/decision-maker and the interaction between AI and the environment.</p>

### 1.8 Societal concerns

<i>Societal concerns</i>
<b>Description</b>
<p><b>Responsibility:</b> Provide the capacity to evaluate the quality of the AI decisions and their corresponding impacts in case of low-quality decisions. Provide mitigation mechanisms to ensure the security, integrity, validity, and accuracy of the AI assistant.</p> <p><b>Explainability and transparency:</b> Disclose to stakeholders the evaluation methods used to assess robustness, explain AI failures (e.g., the impact of input data contamination, communications failure), and allow them to submit test cases and adversarial examples.</p> <p><b>Accountability:</b> Mitigate, detect, and correct erroneous or harmful AI decisions when operating the model.</p> <p><b>Safety and security:</b> The AI system should perform consistently across different scenarios and consider the complexity of the environment in which the AI system will be used. The key question is to understand if technology is fit for its purpose and real-world operating conditions.</p>
<b>Sustainable Development Goals (SGD) to be achieved</b>
<p>SGD7. Affordable and clean energy / SGD13. Climate action</p>

## 2 Environment characteristics

See UC1.Power Grid.

## 3 Technical details

### 3.1 Actors

<i>Actor Name</i>	<i>Actor Description</i>
AI assistant	AI agents provide assistance to human operators. It takes information from the environment to search for recommendations and aid the human operator. In the training phase, it can act on the environment to evaluate its recommendations. In the evaluation/testing phase, the actions on the environment should be performed by the human operator only.
Human operator	A member of TSO's team is in charge of monitoring the grid and taking action on the environment (see "stakeholders" paragraph).
Environment	The human operator will interact with the Digital Environment (illustrated in the Figure below) and the AI assistant through an interface. It can be a digital environment, which is a digital model of the transmission grid, which includes unplanned events that are modeled as events appearing in predefined moments (defined directly in time series). In a real-world implementation, it is the physical environment.

### 3.2 References of use case

<i>References</i>						
<i>No.</i>	<i>Type</i>	<i>Reference</i>	<i>Status</i>	<i>Impact on use case</i>	<i>Originator / organisation</i>	<i>Link</i>
1	AI competition	Paris Region AI Challenge for Energy Transition, Low-carbon Grid Operations, April 2023	Public	The track "Sim2Real" has inspired the use case	Paris Region, RTE	Paris Region <sup>33</sup>

<sup>33</sup><https://www.iledefrance.fr/toutes-les-actualites/entreprises-et-chercheurs-participez-au-challenge-ia-pour-la-transition-energetique>

## 4 Step-by-step analysis of use case

### 4.1 Overview of scenarios

All scenarios happen in an intraday time horizon, meaning not more than a 24-hour forecast period. Scenario 2 is a “nice to have” scenario, which means that it is of less priority than the other scenarios for the project.

Notes regarding scenario and environment data:

- Scenario #1 uses Power Grid Assistant data from Use Case 1, scenario 1, which is progressively altered (e.g., replace data points by zero or NaN if possible). For harder cases, the following modifications could be:
  - a. a grid element is added or removed on the zone
  - b. generation changes (e.g., increase of RES generation capacity)
  - c. the AI assistant is used in a different zone
- It is specific to scenario #2

Note regarding requirements: The column “requirement” for the scenarios’ steps has been left empty for the moment. That column will get more relevant in later stages of the integration/development when moving for a field demonstration or to demonstrate a technology with higher maturity.

Scenario conditions					
No.	Scenario name	Scenario description	Triggering event	Pre-condition	Post-condition
1	Adaptation to real-world conditions	<p>The AI assistant’s robustness is tested on bad or low data. The situation can worsen to the point where the transmission grid state can’t be estimated properly by the AI assistant, which can’t propose any action recommendation.</p> <p><i>Note: other more difficult cases could be:</i></p> <ul style="list-style-type: none"> <li>• new grid elements on the zone</li> <li>• the AI assistant is used on a different transmission grid than in the training phase (transfer learning)</li> </ul>	<p>Issues and inconsistencies are present in the data, and data are also missing.</p> <p>Forecasting of transmission grid state is challenged or can’t even be performed by the AI assistant because the quality of input data is too low and/or the proportion of missing data is too high.</p>	Run scenario 1 from the use case Power Grid Assistant	<p>The recommendations from the AI assistant make the human operator aware of the sensitivity to the uncertainty of recommended actions.</p> <p>All new episodes are rerun with an AI assistant trained on the episodes with altered perception. The result is compared with AI assistants not trained in these conditions.</p>

Scenario conditions					
No.	Scenario name	Scenario description	Triggering event	Pre-condition	Post-condition
2	(Nice-to-have scenario) Additional information from the human operator	<p>The effect of actions recommended by the AI assistant is challenged by unexpected events or dynamics, like the shift of distribution (in this scenario, RES generation). Due to the magnitude of change, specific information is needed from the human operator.</p> <p><i>Note 1: a subcase could be added where the human operator is not able to provide information to the AI assistant.</i></p> <p><i>Note 2: other cases could be where one or several prerequisites (e.g., data) needed to perform an action recommended by the AI assistant are missing or have changed.</i></p> <p><i>Note 3: This scenario shares a lot in common with the first scenario of the use case Power Grid Assistant. However, even if it also includes an unplanned event, the one considered is a shift of the distribution of RES generation pattern, which is not an event monitored in the same way as a list of predefined outages. In addition, this scenario also includes the use of additional information from human operators by the AI assistant.</i></p>	<p>The AI assistant has provided one or several action recommendations.</p> <p>The human operator has assessed that the proposed actions are not feasible or didn't have the expected consequences for the transmission grid's state.</p>	<p>The real-time behavior of the transmission grid is significantly different from the simulation.</p>	<p>The AI assistant proposes new alternative actions with the help of information provided by the human operator.</p>

## 4.2 Steps of scenario 1

Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged	Requirement
1	Start	Run episodes from scenario 1 from the use case Power Grid Assistant	<p>AI assistant's perception of the environment is altered.</p> <p><i>Harder cases could be:</i></p> <ul style="list-style-type: none"> <li>• a grid element is added or removed,</li> <li>• the AI assistant is used in a different zone from the one used in the training</li> </ul>	(empty)	(empty)	(empty)	(empty)
2	Action recommendations	The human operator processes the recommendations	<p>The AI assistant proposes action recommendations to the operator</p> <p>The recommendations from the AI assistant make the human operator aware of the sensitivity to the uncertainty of recommended actions.</p>	AI assistant	Human operator	AIR	(left empty)
3	Unfeasibility of action recommendation	The AI assistant can't provide recommendations	The AI assistant can't propose action recommendations to the operator and indicate the reasons.	AI assistant	Human operator	AIAL	(left empty)

Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged	Requirement
4	Evaluation	The AI assistant's handling of the real world is evaluated	<p>All episodes are rerun with an AI assistant trained on the episodes with altered perception. The result is compared with AI assistants not trained in these conditions to evaluate especially what will be the reaction of the human operator when working with each assistant.</p> <p><i>Note: the following distinction shall be between:</i></p> <ul style="list-style-type: none"> <li><i>False Positives: AI assistant doesn't raise inaccuracy alert, but it can't properly assess the situation,</i></li> <li><i>False Negatives: The AI assistant indicates it has insufficient data to estimate the state of the grid, but it does have the required data.</i></li> </ul>	(empty)	(empty)	(empty)	(left empty)

### 4.3 Steps of scenario 2

For each step, an example of operational business context is given; this will be further detailed during the definition of scenario data. Here, the scenario starts when handling a planned maintenance operation on the grid at the beginning of an operator's shift (start as scenario 1 from the Power Grid Assistant UC).

Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged	Requirement
1	Start	The human operator prepares his/her shift	<b>Example of context:</b> At 08:00 AM, the previous operator ended his/her shift. A planned outage of line L0 starts at 09:00 AM. For this planned outage, a load limitation has been agreed upon beforehand with DSO on a selected set of substations (100 MW max load), knowing that this load is netted with connected RES generation.	(empty)	(empty)	(empty)	(empty)
2	Operator's action	The operator prepares planned outage	<b>Example of context:</b> The operator calls the DSO to confirm that the limitation is implemented before the beginning of the outage The operator fully disconnects line L0 The operator confirms to the maintenance team that the maintenance work can start.	Human operator	Environment	HA	(left empty)
3	Overload forecasted	The AI assistant raises an alert	<b>Example of context:</b> A potential overload is foreseen (N situation) starting at 12:00 PM on the line L1. This overload, if confirmed, needs remedial action (else operational limits would be violated)	AI assistant	Human operator	AIAL	(left empty)
4	Action recommendations	The human operator processes the recommendations	<b>Example of context:</b> AI assistant proposes one possible curative remedial action (the same as the one foreseen during operational planning preparation of the outage): it consists of opening a line L2	AI assistant	Human operator	AIR	(left empty)



Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged	Requirement
5	Overload alert	The AI assistant raises an alert	<b>Example of context:</b> The flow on line L1 is increasing and exceeds the admissible flow in the N situation	AI assistant	Human operator	AIAL	(left empty)
6	Operator's decision	The operator decides to implement an action	<b>Example of context:</b> The human operator decides to perform the recommended action and opens the line L2, which brings the flow on line L1 back to admissible level	Human operator	AI assistant	D	(left empty)
7	Unplanned event	Change of forecasted flows	<b>Example of context:</b> The flow on the line is different from what is forecasted  This can correspond to real situations, e.g., sudden gusts of wind (or, on the contrary, sudden drops)	Environment	AI assistant	E	(left empty)
8	Overload alert	The AI assistant raises an alert	<b>Example of context:</b> The flow on the line L1 exceeds again the admissible flow in N situation	AI assistant	Human operator	AIAL	(left empty)
9	Action recommendations	The human operator processes the recommendations	<b>Example of context:</b> AI assistant proposes only one possible curative remedial action: load shedding	AI assistant	Human operator	AIR	(left empty)
10	New information from human operator to AI Assistant	The human operator provides additional information in the context of constraint-solving	<i>Note: a subcase could be added where the human operator is not able to provide information to the AI assistant</i>  <b>Example of context:</b> After analysis, the human operator realizes that the load on the agreed substations exceeds the agreed volume of 100 MW The human operator checks with the DSO that one transformer can be opened (no risk of load shedding) and adds this as a possible remedial action in the AI assistant	Human operator	AI assistant	NINF	(left empty)

<i>Step no.</i>	<i>Event</i>	<i>Name of process/ activity</i>	<i>Description of process/ activity Service</i>	<i>Information producer (actor)</i>	<i>Information receiver (actor)</i>	<i>Information Exchanged</i>	<i>Requirement</i>
11	Action recommendations	The human operator processes the recommendations	<b>Example of context:</b> The AI assistant assesses possible actions and recommends going for opening the transformers	AI assistant	Human operator	AIR	(left empty)
12	Operator's decision	The operator decides to implement an action	<b>Example of context:</b> The human operator decides to perform the recommended action	Human operator	AI assistant	D	(left empty)
13	Operator's action	The operator implements an action	<b>Example of context:</b> The operator opens the transformer, which brings the flow on line L1 back to admissible level	Human operator	Environment	HA	(left empty)

## 5 Information exchanged

<b>Information exchanged (ID)</b>	<b>Name of information</b>	<b>Description of information exchanged</b>
HA	Action implemented by a human operator	Action (e.g., topology) implemented by human operator
AIAL	AI assistant alert	AI assistant alert for an overload occurring on one or several grid elements. AI assistant alert for reached time limit of a given action
AIR	AI assistant recommendations	List of remedial action recommended by the AI assistant
D	The decision from a human operator	Human operator's choice
E	Environment information	Information on the environment, e.g., outages. <i>In case an adversarial agent is used to model unplanned events, this information would be replaced by an "adversarial attack".</i>
NINF	New information	Information related to the environment context that is not known by the AI assistant

## 6 Requirements

<b>Requirements</b>		
<b>Categories ID</b>	<b>Category name for requirements</b>	<b>Category description</b>
Ro	Robustness	It encompasses both its technical robustness (the ability of a system to maintain its level of performance under a variety of circumstances) and its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no unintentional harm can occur. <i>Source: EU-U.S. Terminology and Taxonomy for Artificial Intelligence. First Edition</i>
E	Efficiency	The ability of an AI system to achieve its goals or perform its tasks with optimal use of resources, including time, computational power, and data.
I	Interpretability	Make the behavior and predictions of AI systems understandable to humans, i.e., the degree to which a human can understand the cause of a decision. <i>Source: Molnar, Christoph. Interpretable machine learning. Lulu.com, 2020.</i>
FAIR	Non-discrimination and fairness	This means that AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality, and cultural diversity while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law. <i>Source: EU AI Act</i>

HAO	Human Agency and Oversight	The design phase involves including mechanisms for human intervention and ensuring that people can easily understand and monitor AI systems. During deployment, it means continuous monitoring and evaluation to ensure that the systems act within their ethical boundaries.
Requirement R-ID	Requirement name	Requirement description
Ro-1	Adaption to increased uncertainty	The AI system should demonstrate the ability to sustain operational stability and decision performance in diverse and partially unpredictable scenarios, such as increased forecasting errors, missing data, unavailable control actions, and delayed measurements.
Ro-2	Network change responsiveness	The AI system must be able to handle changes within the transmission grid infrastructure, such as introducing new grid elements and modifying the grid topology as the electrical grid evolves.
Ro-3	Cognitive load and stress	The AI system shall not increase the complexity of the situation and the associated level of stress for human operators (due to additional misinformation).
Ro-4	Reproducibility of recommendations for <i>post-mortem</i> analysis	All recommendations made by the AI system must be reproducible at a later point, given the same input or specific context/conditions. While the actions recommended by the system do not need to be identical in a strict mathematical sense - acknowledging the variability inherent in distributed computing environments - they should be closely aligned and functionally equivalent, ensuring reliable and predictable outcomes under similar conditions. Moreover, it should be possible to trace back which AI model or rules led to the decision(s) or recommendation(s) of the AI system, which is very relevant for audits from the Energy Regulator.
Ro-5	Increase technical robustness to missing or erroneous input data	The training of the AI system should include scenarios with natural or adversarial perturbations in its input/state vector, which can originate from missing or erroneous values from the environment ( <i>or adversarial attacks from agents</i> ).
Ro-6	Robustness to attacks targeting model space and reward function	Reward functions and models should be stored and operated in highly cyber-secure Information Technology systems. In the event of an attack, the previously trained model could be quickly restored. Model training should be done in a secure and controlled digital environment, and model retraining is possible.
E-1	Computational efficiency	The AI system must be designed to ensure efficient training and inference capabilities on various computer hardware, from small-scale development setups with limited processing power to configurations involving multiple servers and GPUs.

I-1	Adaptability to different levels of interaction and human operator preferences	Each operator has its own preferences (e.g., one operator can be more risk averse than others): ideally, the AI assistant interacting with a given operator could provide decision support that fits the preferences of this operator but is not necessary of another, especially given the type of situation that can require more attention. Thus, the AI system shall be able to interact with the human operator according to his/her preferences and experience, such as a) fully manual, b) get notified every time an overload is detected, and c) only get notified when the AI assistant is not confident enough.
FAIR-1	Avoid creating or reinforcing unfair bias in the AI system	The system must not unfairly favor specific producers or consumers of electrical energy. A level playing field in the electricity market, as well as fair competition, must be provisioned. Measures must be implemented to ensure these fairness constraints are observed.  <i>Note that:</i> 1) <i>Occurring bias may very well originate from technical or physical limitations of electrical grid operations and hence may (in part or wholly) not be avoidable.</i> 2) <i>Requiring the AI system to adhere to fairness standards that are not required from existing alternative techniques may put it at a disadvantage, especially if those originate from the source of the previous issue.</i>
FAIR-2	Regular monitoring of fairness	Using the physical equations of the power grid, it is possible to compare the decisions made by the AI system and the impact that other grid users would have in solving the technical problem. For instance, <i>ex-post</i> , it is possible to run an optimal power flow with the redispatch costs and compare its solution with the AI system. Having a least-cost solution is the primary goal. Metrics such as Jain's fairness index have been used to evaluate fairness in load shedding <sup>34</sup> and fairness in renewables' curtailment <sup>35</sup> .
HAO-1	Additional training about AI for human operators	The type of recommendation from this use case is already known by the human (i.e., the same as traditional tools in power system control rooms), but humans should be trained to understand the rationale behind the AI system (e.g., understand how reinforcement learning works) and its limitations.

## 7 Common Terms and Definitions

Common Terms and Definitions	
Term	Definition
TSO – Transmission System Operator	A natural or legal person is responsible for operating, ensuring the maintenance of, and, if necessary, developing the transmission system in a given area and, where applicable, its interconnections with other systems and for ensuring the long-term ability of the system to meet reasonable demands for the transmission of electricity. Source: Directive 2009/72/EC and ENTSOE glossary

<sup>34</sup> F. Moret and P. Pinson, "Energy Collectives: A Community and Fairness Based Approach to Future Electricity Markets," IEEE Trans. Power Syst., vol. 34, no. 5, pp. 3994–4004, Sep. 2019.

<sup>35</sup> M. Z. Liu Liu, A. T. Procopiou, K. Petrou, L. F. Ochoa, T. Langstaff, J. Harding, and J. Theunissen, "On the Fairness of PV Curtailment Schemes in Residential Distribution Networks," IEEE Trans. Smart Grid, vol. 11, no. 5, pp. 4502–4512, 2020.

<b>Common Terms and Definitions</b>	
<b>Term</b>	<b>Definition</b>
EMS – Energy Management System	Optimal control center solution to enable secure, efficient, and optimized operation of the electric power system.
Contingency (“N-1” situation)	Electric system’s state after the loss of one grid element, and possibly several grid elements, depending on the TSO’s policy
Load (or power) flow calculation	Calculations are used to determine the voltage, current, and real and reactive power at various points in a power system under steady-state conditions.

## UC1.RAILWAY: AUTOMATED RE-SCHEDULING IN RAILWAY OPERATIONS

### 1 Description of the use case

#### 1.1 Name of the use case

ID	Application Domain(s)	Name of Use Case
UC01.Railway	Railway network	Automated re-scheduling in railway operations

#### 1.2 Version management

Version Management			
Version No.	Date	Name of Author(s)	Changes
0.1	12.04.2024	Roman Ließner, Irene Sturm, Adrian Egli	Initial Version (import from UC1.Railway short)
0.2	14.04.2024	Manuel Renold, Adrian Egli	Checked alignment use cases/framework and more update
0.3	16.04.2024	Ricardo Bessa	Revision
0.4	17.04.2024	Julia Usher	Revision
0.5	25.04.2024	Adrian Egli, Daniel Boos, Irene Sturm, Roman Ließner, Manuel Schneider	Final Revision
0.6	30.05.2024	Adrian Egli	Revision: Action space
1.0	08.07.2024	Ricardo Bessa	Final version

#### 1.3 Scope and objectives of use case

Scope and Objectives of Use Case	
<b>Scope</b>	<p>Traffic density on the European rail networks is constantly increasing. This increases the complexity of rail traffic management in operations: timetables are constructed to maximize utilization of the network's capacity. At the same time, new construction or maintenance of railway infrastructure must be planned and carried out efficiently.</p> <p>In railway operations, the already densely planned schedules are disturbed by unexpected events, such as delays, infrastructure defects, or short-term maintenance. The execution of the planned timetable can only be achieved by acting on these events with frequent adaptation and re-scheduling of the planned train runs. Today, maintaining smoothly running operations requires that in operational centers, highly skilled personnel monitor the flow of traffic day and night, and quickly make re-scheduling decisions.</p>
<b>Objective(s)</b>	The system's objective is to fully automate re-scheduling in railway operations to fulfill all offered services and minimize delays for the customer (passenger).
<b>Deployment model</b>	Cloud services and on-premises.

#### 1.4 Narrative of use case

Narrative of Use Case
-----------------------

### Short description

In railway operations, traffic on the network is planned to fulfill the intended service that was contracted with the Railway Undertaking Operating Managers (RUOM), e.g., to execute train runs on the network so that the requested commercial stops are fulfilled in a punctual manner. In operations, such a pre-planned schedule is executed.

Unexpected events, such as infrastructure malfunctions or delays occur. In case of such a deviation, the automated system must re-calculate the schedule so that the requested services can be fulfilled with as little delay as possible. Adapting the schedule includes interventions, such as changing the speed curves of trains, changing the order of trains at the infrastructure element, changing the routes of trains, or changing the platform of a commercial stop in a station. A highly automated AI-based system is designed to manage and optimize railway schedules in real time, ensuring efficient rail network use while minimizing delays for passengers. The system is constantly monitored by a human operator who can adjust the system's configuration and identify the need for adaptation and re-training.

### Complete description

**Description of the re-scheduling task:** Re-scheduling trains in railway operations means monitoring the movement of trains on a railway network and reacting to unexpected events, such as signal failures, track blockages, weather events that disrupt operations, or other significant delays, and also proactively to predicted deviations that affect planned operations in the future. Re-scheduling measures include changing a train's speed, path, or platform. In a densely utilized railway network, local re-scheduling decisions potentially affect the entire flow of traffic, and their effect can propagate far into the future. This means that the re-scheduling task is a very complex decision-making task that must integrate a lot of context information under time constraints

**System description and role of the human operator:** An AI-based re-scheduling system performs the re-scheduling task in a highly automated manner. This system observes the real-time state of all the trains and tracks in the control area of interest and automatically detects the need to intervene, decides on an intervention, and executes this intervention. Such an AI system for highly automated re-scheduling in operations is something new and unusual. The approach followed here can be understood as a first step towards introducing such a system. The highly automated AI system is treated as a new tool that is supervised and evaluated by an expert. The goal is to find the limits of the automated system as a starting point for improving and configuring it.

In operations, the AI system re-schedules in a fully automated manner while the human supervisor monitors:

- The system's state in operations (e.g., number of trains, potential bottleneck in current and planned network usage)
- KPIs for the actual situations (e.g., current delay)
- Confidence/certainty of the AI system
- Intensity of intervention (how much changes to the current operational plan did the AI perform, e.g., change platform)

The supervisor uses this information to:

- Decide at which point it would be advisable to switch off the AI system and take over control.
- Decide to re-configure/adjust the system in operations.

The overarching goal in this setup is to learn the existing solution's limits: in which situations does the AI system reach appropriate decisions? These insights should not only be generated from metrics extracted in tests and analyzed post-hoc but also in a realistic operational context with which the human operator is familiar.

**Operational scenario:** For an operational scenario, there exists a definition of the intended service that was contracted with the network operator's customers (Railway Undertaking Operating



Managers (RUOMs)), e.g., a set of train runs with a sequence of commercial stops. For all commercial stops, there exists a time constraint, defining:

- Latest arrival
- Minimal dwell time
- Earliest departure

An initial schedule exists that is executable and fulfills the intended services, such as the arrival and departure times of trains at commercial stops, while taking into account operational requirements (safety systems, additional constraints). A schedule contains all the information that is needed to execute train runs.

A schedule is **acceptable** if all hard constraints are fulfilled:

- Commercial stops were performed in the right order before the end of the scenario.
- Minimal dwell time for each stop has been respected.
- Earliest departures for each stop have been respected.

A schedule is punctual, i.e., fully fulfills the intended service; if the schedule is acceptable for all commercial stops, the constraint of “latest arrival” has been respected.

The following steps are performed in the use case:

1. **Definition of System Parameters:** Detailed parameters are set for the pre-planned schedule, including the prioritization of trains in case of disruptions, acceptable delay margins, and specific criteria for train prioritization (e.g., passenger load and destination importance). This step also includes the configuration of safety systems, network capacity limits, and any special operational requirements unique to certain routes or times.
2. **Schedule Execution:** The initial operational plan is executed in operations. This includes the deployment of trains according to the pre-planned schedule, monitoring of train movements, adherence to the sequence of commercial stops, and ensuring compliance with operational requirements like safety systems. The state of the system is also displayed to the human supervisor in an appropriate manner.
3. **Triggering Re-scheduling:** The re-scheduling process can be initiated by a variety of triggers, such as infrastructure changes (e.g., blocked tracks, malfunctioning switches), train delays, or equipment malfunctions. The system is designed to detect these deviations in real time and assess their impact on the overall schedule. The exact nature of this trigger or several different triggers needs to be defined and should also be configurable for usage.
4. **Display of Deviation and Triggering Re-calculation:** Upon detecting a deviation, the system provides a detailed display of the issue, including its nature, location, and expected impact on the schedule. It then notifies the human supervisor and initiates the re-calculation process.
5. **Automated Schedule Re-calculation:** The Traffic Management System (TMS) automatically recalculates the schedule from the point of deviation to the end of the operational scenario. The goal is to create an adapted schedule that is acceptable (meeting all hard constraints) and minimizes total delays, particularly focusing on the 'latest arrival' times at commercial stops.
6. **Execution of Adapted Schedule:** The newly adapted schedule is then put into operation. The system continuously monitors for any further deviations and adjusts the schedule as needed to maintain operational efficiency and adherence to time constraints.

**Human Review and System Adjustment:** A human supervisor reviews the performance of the system, analyzing how effectively it responded to deviations and the impact on service delivery. Based on this review, adjustments are made to the system's parameters, such as altering the prioritization criteria, adjusting acceptable delay thresholds, or refining the algorithm for schedule recalculations. This step ensures continuous learning and improvement of the system based on operational experiences and organizational goals.

#### **Stakeholders**

**Railway network operator:** Operator of the railway network in charge of maintaining the flow of traffic on the railway network to provide high quality-of-service to their direct customers (RUOMs) and the passengers.

**Network supervisor:** Human supervisor of the automated railway system (something like the former dispatcher who is not dispatching himself anymore but monitoring the system state),

**RUOM:** Railway Undertaking Operation Manager offering passenger and freight traffic services.

**Neighboring areas of control/operational centers.**

**Passenger:** The primary end-user of the railway services whose travel experience and satisfaction are directly impacted by the efficiency and punctuality of train operations.

**Government and society:** The quality of railway services is a concern of the government and society.

**Stakeholders' assets, values**

**Railway network operator:**

- Available capacity on the network: a low-quality re-scheduling functionality will consume more capacity on the network.
- Reputation: low performance of the AI system can lead to a bad reputation in terms of operational stability, punctuality, etc., which might cause customers to not rely on and to use less the services offered. This also concerns network operators, RUOM, and passengers.
- Legal and regulatory framework: Regulations with the discrimination-free treatment of RUOMs.
- Unintended behavior of the AI system and actions by malicious actors can potentially compromise the safety of the train passengers, personnel on the train, and on and in proximity to the tracks, as well as infrastructure like tracks, power lines, tunnels, stations, etc.

**Human dispatcher:**

- Damage to the reputation, safety issues as well as a potential general perception of an opaque AI-system being in control of running trains can cause a decrease in the trustworthiness of the railway operator from a customer perspective, both for individual travelers and cargo transport.

The usefulness and understandability of the AI-system output to the dispatcher may influence the trustworthiness of the AI-system from the perspective of the dispatcher. Low trustworthiness might render the use of the AI system irrelevant as the dispatcher will not trust the options generated by the system, and the assumed benefit will not materialize.

**System's threats and vulnerabilities**

**Accountability:** who is responsible for delays and, in general, bad performance of the AI system.

**Security:** A highly automated AI system introduces the risk of severe abnormal situations on the railway network. Although in railway systems, the immediate danger of train collision is addressed by separate systems that the AI system will not control, there is a risk of severe traffic congestion with significant economic effects on the network in case of a malfunctioning AI.

1.5 Key performance indicators (KPI)

Name	Description	Reference to the mentioned use case objectives
Acceptance score	Tracks the frequency of human operator interventions in AI decisions. Target: Reduce to less than x% of cases. Calculation: (Number of human interventions / Total AI decision instances) x 100.	Reflects the reliability and trust of the AI system.
Punctuality	Measures the percentage of trains arriving at their destinations on time. Target: Achieve a punctuality rate of x% or higher. Calculation: (Number of on-time arrivals / Total number of arrivals) x 100.	Linked to the objective of minimizing delays.
Response time	Assesses the speed at which the AI system responds to disruptions or changes. Target: Response within x minutes of disruption detection. Calculation: Average time taken from disruption detection to system response.	Related to the objective of rapid re-scheduling.
Delay Reduction Efficiency	Quantifies the effectiveness of the system in reducing delays. Target: Reduce overall delays by 30%. Calculation: (Total delay duration before AI implementation - Total delay duration after AI implementation) / Total delay duration before AI implementation.	Linked to the objective of minimizing delays.
Trust towards the AI-System	<p><i>“(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)”</i> (Cahour &amp; Forzy, 2009, p. 1261).</p> <p>The human operators' trust in the AI tool can be measured using the Scale for XAI (Hoffman et al., 2018) or similar.</p>	Linked to the human operator's appropriate trust in the AI system as a necessary precondition of adequate use.
Human motivation	<p><i>“Intrinsic motivation is defined as the doing of 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”</i> (Ryan &amp; Deci, 2000, p. 54).</p> <p>The human operators perceived internal work motivation can be measured by using the Job Diagnostic Survey (Hackman &amp; Oldham, 1974) or similar. The questionnaire needs to be adapted to the AI context (e.g., problem detection with AI assistance).</p>	This is linked to the necessary motivation of the human operator to use the AI for complete a task and reach corresponding objectives.

<p>Human control/autonomy over the process</p>	<p>Autonomy is the degree to which the job provides substantial freedom, independence, and discretion to the employee in scheduling the work and in determining the procedures to be used in carrying it out” (Hackman &amp; Oldham, 1975, p. 162). It consists of three interrelated aspects centered on freedom in decision-making, work methods, and work scheduling (Morgeson &amp; Humphrey, 2006). Parker and Grote (2022) view job autonomy interchangeably with job control.</p> <p>The human operator's perceived autonomy over the process can be measured by using the Work Design Questionnaire (Morgeson &amp; Humphrey, 2006) or similar. The questionnaire needs to be adapted to the AI context (e.g., problem detection with AI-assistance).</p>	<p>Linked to the perceived control of the human operator as a necessary prerequisite for taking responsibility for the efficiency and effectiveness of one's own work.</p>
<p>Human learning</p>	<p>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, as well as the continuity over time (Alexander et al., 2009).</p> <p>The human operators perceived learning opportunities working with the AI-based system can be measured by using the task-based workplace learning scale (Nikolova et al., 2014) or similar. The questionnaire needs to be adapted to the AI context.</p>	<p>Linked to the objective of mutual co-learning to assist the human operator in improving his/her performance.</p>
<p>Decision support for the human operator</p>	<p>Decision support tools should be aligned with the cognitive decision-making process that people use when making judgments and decisions in the real world and ensure that the human operator retains agency (Miller, 2023). AI decision support tools should, therefore, help people to remain actively involved in the decision-making process (e.g., by helping them critique their own ideas) (Miller, 2023).</p> <p>The decision support for the human operator can be measured based on the criteria for good decision support (Miller, 2023) or similar. The instrument needs to be further developed.</p>	<p>Linked to the appropriateness of AI-based support of the human operator's decision-making process.</p>
<p>Ability to anticipate</p>	<p><i>“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”</i> (Hollnagel, 2015, p. 4).</p> <p>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 an AI-based assistant can be measured by using the Rigor-</p>	<p>Linked to AI-based enabling of human operators to minimize delays for the customers.</p>

	Metric for Sensemaking (Zelik et al., 2010) or similar. The instrument needs to be further developed and adapted to the AI context.	
Situation awareness	<p>“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, p. 12).</p> <p>The human operator’s situation awareness can be measured by using the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1988) or similar.</p>	Linked to the AI-based assistance of the human operator for developing an appropriate situation awareness.

### 1.6 Features of use case

<b>Task(s)</b>	Planning, prediction, optimization, interactivity, and recommendation.
<b>Method(s)</b>	Reinforcement learning has been applied to this use case, but other AI approaches are possible.
<b>Platform</b>	<a href="#">Flatland</a> digital environment.

### 1.7 Standardization opportunities and requirements

<i>Classification Information</i>
<b>Relation to existing standards</b>
<p><b>ISO/IEC 23894:2023, Information technology — Artificial intelligence — Guidance on risk management.</b> Autonomous management and optimization of railway scheduling in real-time are high-stakes tasks, and therefore, risk management specifically related to AI is fundamental.</p> <p><b>ISO/IEC 38507:2022, Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations.</b> Autonomous AI requires an analysis of governance implications and also a redefinition of the organization structure.</p> <p><b>ISO/IEC 24029-2:2023, Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for using formal methods.</b> Since artificial neural networks can be a component of the autonomous AI system, formal methods to assess the robustness properties of neural networks are fundamental to certify and monitor autonomous systems.</p> <p>In railway transport, there are different levels of automation (Grade of Automation, GoA) defined in the <b>IEC 62267</b> Standard ("Railway applications - Automated urban guided transport (AUGT) - Safety requirements"). This standard covers high-level safety requirements applicable to automated urban guided transport systems, with driverless or unattended self-propelled trains, operating on an exclusive guideway.</p> <p><b>DIN EN 50126, Railway Applications – The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS).</b> It considers the generic aspects of the RAMS life cycle and provides a description of a Safety Management Process. It provides guidelines for defining requirements, conducting analyses, and demonstrating the reliability, availability, maintainability, and safety aspects throughout the lifecycle of railway applications.</p> <p><b>DIN EN 50128, Railway applications – Communication, signaling and processing systems.</b> Outlines the procedural and technical criteria for crafting software intended for programmable electronic systems in railway control and protection applications.</p>
<b>Standardization requirements</b>
<p>Opportunities for standardization and deriving recommendations for critical operations management and support, especially regarding co-decision-making and human-computer interaction, as well as safety requirements. See also UC2.Railway.</p>

### 1.8 Societal concerns

<i>Societal concerns</i>
<b>Description</b>
<p><b>Privacy and data protection:</b> The use of AI in railway scheduling involves the collection and analysis of large volumes of data, including potentially sensitive information. There is a concern about how this data is stored, processed, and protected, especially in compliance with data protection regulations like GDPR. Ensuring the privacy and security of passenger and employee data is paramount.</p> <p><b>Transparency and accountability:</b> There is a societal demand for transparency in how AI systems make decisions, especially in critical infrastructure like railway systems. The public might be</p>

<p>concerned about the lack of understanding of AI decision-making processes and the accountability mechanisms in place in case of failures or errors.</p> <p><b>Employment and skill shift:</b> The automation of train scheduling might lead to concerns about job displacement and the need for reskilling of railway staff. While AI can optimize operations, it also changes the nature of work, requiring a shift in skills for human operators who now need to oversee and interact with advanced AI systems.</p> <p><b>Public trust and acceptance:</b> For the successful implementation of AI in public transportation, gaining and maintaining public trust is crucial. There may be apprehensions and resistance from the public regarding the shift to AI-driven systems, especially among those accustomed to traditional methods.</p> <p><b>Safety and security:</b> The use of AI systems for critical operational scenarios raises concerns regarding the continued safety and security of these systems. Understanding failure modes, developing robust models, and ensuring resilience to adversarial attacks are among the many topics to be tackled.</p> <p><b>Inequality:</b> Such systems might introduce inequality in service quality for different geographic regions or categories of passengers due to the opacity of the system, bias, and self-learning aspects.</p>
<p><b>Sustainable Development Goals (SGD) to be achieved</b></p>
<p>SDG9. Decent work and economic growth / SDG9. Industry, innovation and infrastructure / SDG11. Sustainable cities and communities / SDG13. Climate action</p>

## 2 Environment characteristics

<i>Characteristics</i>	
<i>Observation space</i>	<p>Partially observable with limitations due to the unpredictable duration of delays and malfunctions.</p> <p>Data update is near real-time (rather seconds than hours).</p> <p>Domain: defined on a continuous space.</p> <p>Size: Depending on the type of observation considered local or global, the total size can depend, but it will generally be very large.</p> <p>Noise: The observation can be noisy due to the communication system and the various signaling devices (<i>signal box</i>).</p> <p>(In addition to more than 10,000 trains (per day), there are over 32,000 signals and over 14,000 switches in the Swiss rail network. All of this information must be considered and observed; thus, the global observation is very large.)</p>
<i>Action space</i>	<p>Mixed action space: actions like which route to take on a switch are discrete, as well as decisions like whether a train should accelerate or decelerate. However, dependent on the algorithmic approach, the rate of acceleration, deceleration, velocity to move forward, and similar can be modeled both discrete and continuous.</p> <p>Size: Depends on the algorithmic approach. While the action space grows linearly with the number of trains for the algorithmic part, it grows exponentially if there is a central actor controlling all the trains. The action space of the human dispatcher is, in any case, exponentially growing with the number of trains. Furthermore, the dimensionality of the action space depends on infrastructure and timetable elements like switches, signals, and scheduled stops. Hereby, the impact on the dimensionality of the action space depends not only on the actor's nature in the algorithmic part but also on the type of task, i.e., if the task is tackled episodically or sequentially on the algorithmic side. For the human dispatcher, the task is generally considered to be sequential since an action is usually dependent on previous actions taken.</p> <p>Time horizon: An action typically takes from a few minutes to a couple of hours.</p> <p>The action space of the flatland environment is 5 (go left, go forward, go right, stop, none). However, each train run (agent) must perform one of these basic actions at</p>

	each decision point (time step). This means that the total number of actions to be selected is very large and stays in linear relation to the number of agents - i.e., in a problem-solving scenario with $n$ agents and $m$ time steps, the actions should be chosen in such a way that the combination of selected actions leads to the desired outcome or optimal solution. Each agent has a set of actions to choose from, from which they must select one at each time step. Therefore, the solution involves $n \times m \times a$ possible actions. (Up to 800 trains run simultaneously on the Swiss rail network. In many cases, they interact directly or indirectly with each other.).
<b>Type of task</b>	The nature of the task depends on the algorithmic approach. While AI models can determine which action to take fully based on the current state without including information about past actions and would therefore be considered episodic, other approaches can, to a large degree, approach problem-solving as a sequential task, for example, if planning is involved. The human dispatcher usually approaches the task sequentially.
<b>Sources of uncertainty</b>	Stochastic, with the following sources of uncertainty: 1) Weather conditions can impact, e.g., the friction of wheels on rails, which leads to different acceleration and deceleration behavior. 2) The travel demand influences both the total load of a train and the delay to board other passengers. 3) Disruptions: Train level – locomotives or another rolling stock issue that may arise and result in a delay; Infrastructure level – signal malfunctions or construction sites. 4) Sensors and communication level – a failure may introduce noise and uncertainty in observing the environment.
<b>Environment model availability</b>	A specific model of the environment is not available. Although a good approximation of it can be achieved as the basic laws of physics are defined and clear. However, a model of the environment will be simplified in general and subject to uncertainty (see above).
<b>Human-AI interaction</b>	Co-learning between the human and AI: The interaction between humans and AI is done just after fully automated rescheduling when the super users analyze the outcome of the operations. (Learning from post-perspective analytics).

### 3 Technical details

#### 3.1 Actors

<b>Actor Name</b>	<b>Actor Description</b>
Dispatcher	The dispatcher is a human responsible for monitoring and analyzing railway traffic. The main role is to ensure the safe and efficient movement of trains by controlling the flow of traffic and making decisions based on real-time information. The dispatcher determines the order of trains and may deviate from planned routes when necessary to accommodate unexpected situations or optimize the overall operation. The decisions play a crucial role in maintaining the smooth functioning of the railway system.
Traffic control system	The traffic control system collects information such as traffic signals, train positions, and current train speeds and also provides a human-machine interface for controlling ongoing traffic. The system's goal is to manage the flow of traffic efficiently, centrally, and safely. This necessitates the comprehensive collection of available information to effectively support the decision-making process, which is primarily performed by human dispatchers. Consequently, the traffic control system is vital and should be



	implemented with a human-centered approach unless a fully automated solution is available.
Train run (Driver)	A train run refers to the operation of a train on a specific route or journey from one station to another. It encompasses the entire process of a train traveling along its designated path, including departure from the originating station, intermediate stops (if any), and arrival at the destination station. The current position and speed of the train are communicated to the traffic control system.

## 4 Step-by-step analysis of use case

### 4.1 Overview of scenarios

Scenario conditions					
No.	Scenario name	Scenario description	Triggering event	Pre-condition	Post-condition
1	Re-Scheduling at the occurrence of infrastructure malfunction	The automated railway management system faces a challenge when a sudden infrastructure malfunction occurs (trigger event). This requires an immediate and strategic response to ensure continued service delivery and minimize disruptions.	A change in the infrastructure, e.g., a track becomes unexpectedly blocked	Intended service: a set of train runs with Start- and end locations, a sequence of commercial stops, both with time information (Latest arrival, minimal dwell time, earliest departure). An initial (microscopic) operational plan that is executable and fulfills the intended services, such as the arrival and departure times of trains at commercial stops.	The system has produced a new operation plan that is executable in the simulation and leads to an “acceptable” state at the end of the scenario.
2	Emergency response to weather challenges	This scenario deals with sudden weather challenges, such as extreme weather conditions, impacting railway operations.	A weather challenge arises, such as a severe storm, heavy snowfall, or flooding, affecting parts of the railway network.	A standard operational plan is in place, but it does not account for a general degradation of the state of operations, such as a general reduction of speed in a larger part of the network or the entire network.	The system quickly evaluates the impact of the environmental challenge on the network. It re-calculates a plan that adapts to the new situation.
3	Closure of a large station	This scenario addresses the challenge of adjusting the schedule in case of a closure of a whole station.	Closure of a station.	A standard operational plan is in place that foresees a number of trains performing commercial stops in the affected station.	Re-calculated plan

#### 4.2 Steps of the training scenario

Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged
1	Start	Definition of system parameters	Detailed parameters are set for the pre-planned schedule, including the prioritization of trains in case of disruptions, acceptable delay margins, and specific criteria for train prioritization (e.g., passenger load and destination importance). This step also includes the configuration of safety systems, network capacity limits, and any special operational requirements unique to certain routes or times.	Administrator	Network Operator	SYSPAR
2	System params defined	Schedule Execution	The initial operational plan is executed in operations. This includes the deployment of trains according to the pre-planned schedule, monitoring of train movements, adherence to the sequence of commercial stops, and ensuring compliance with operational requirements like safety systems. The state of the system is also displayed to the human supervisor in an appropriate manner.	Dispatcher	TMS	EXECPLAN
3		Triggering Re-scheduling	The re-scheduling process can be initiated by a variety of triggers defined by the scenarios listed in 4.1. Examples of such triggers are infrastructure changes (scenario 1), heavy weather events (scenario 2) or station closures (scenario 3). The system is designed to detect these deviations in real time and assess their impact on the overall schedule. The exact nature of this trigger or several different triggers needs to be defined and should also be configurable for usage.			

4		Display of Deviation and Triggering Re-calculation	Upon detecting a deviation, the system provides a detailed display of the issue, including its nature, location, and expected impact on the schedule. It then notifies the human supervisor and initiates the re-calculation process.	TMS	Dispatcher	STATE
5		Automated Schedule Re-calculation	The Traffic Management System (TMS) automatically recalculates the schedule from the point of deviation to the end of the operational scenario. The goal is to create an adapted schedule that is acceptable (meeting all hard constraints) and minimizes total delays, particularly focusing on the 'latest arrival' times at commercial stops.	TMS	Dispatcher, Simulation	EXECPLAN
6		Execution of Adapted Schedule	The newly adapted schedule is then put into operation. The system continuously monitors for any further deviations and adjusts the schedule as needed to maintain operational efficiency and adherence to time constraints.			
7		Human Review and System Adjustment:	A human supervisor reviews the performance of the system, analyzing how effectively it responded to deviations and the impact on service delivery. Based on this review, adjustments are made to the system's parameters, such as altering the prioritization criteria, adjusting acceptable delay thresholds, or refining the algorithm for schedule recalculations. This step ensures continuous learning and improvement of the system based on operational experiences and organizational goals.	TMS	Dispatcher	STATE

## 5 Information exchanged

<i>Information exchanged</i>		
<i>Information exchanged (ID)</i>	<i>Name of information</i>	<i>Description of information exchanged</i>
<a href="#">SYSPAR</a>	System Parameters	A series of parameters is necessary to initialize the environment and provide all operative information to the agent(s).
<a href="#">EXECPLAN</a>	Operational plan	The planned schedule is to be executed, including information such as commercial stop sequence and operational requirements.
<a href="#">STATE</a>	State of the system	Detailed information on the current state of the system. Particular focus is given to any information about deviations from the expected system state.

## 6 Requirements

<b>Requirements</b>		
<b>Categories ID</b>	<b>Category name for requirements</b>	<b>Category description</b>
Ro	Robustness	It encompasses both its technical robustness (the ability of a system to maintain its level of performance under a variety of circumstances) as well as its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no unintentional harm can occur. <i>Source: EU-U.S. Terminology and Taxonomy for Artificial Intelligence. First Edition</i>
E	Efficiency	The ability of an AI system to achieve its goals or perform its tasks with optimal use of resources, including time, computational power, and data.
I	Interpretability	Make the behavior and predictions of AI systems understandable to humans, i.e., the degree to which a human can understand the cause of a decision. <i>Source: Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.</i>
Re	Regulatory and legal	The AI system's capacity to meet its objectives while complying with relevant laws, regulations, and ethical standards.
Fa	Fairness	Ensure the recommendations and predictions of the AI system are in line with the principles of fairness (i.e., fair distribution of the benefits and strain/harm)
O	Other	Other non-function requirements related to environmental concerns and maintenance
<b>Requirement R-ID</b>	<b>Requirement name</b>	<b>Requirement description</b>
Ro-1	Reasonable recommendations in new situations (not seen during model training)	Systems provides reasonable solutions for situations not seen during training.

Ro-2	Good performance in operating scenarios with high variability	The system performs well in situations with many fast-changing elements
Ro-3	Retrospective quality control	The quality of provided options can be assessed in retrospect
E-1	Capacity to handle operating scenarios with high complexity	The system derives options fast and with high quality in complex situations with many trains, switches, and other elements involved.
E-2	Scalability	Concerns the system's ability to handle growth, such as increased train traffic or network expansion, without performance degradation. This ensures the system remains effective as the scale of railway operations increases.
E-3	Generalization to different scenarios	The system's ability to handle previously unseen scenarios and generalize to areas of observation and action space not visited during training (e.g., different speed profiles, rails configuration etc.)
Re-1	Compliance with legal standards and regulations	Adherence to data protection laws, safety regulations, cybersecurity, and ethical guidelines governing AI systems in public transportation and the EU AI Act.
I-1	Interpretability of suggestions	The process through which the AI system learns and operates, including how it generates suggestions, is transparent and understandable to the human dispatcher. Further, the decision-making that leads to the suggestion, as well as its limitations, are explained to the human dispatcher.
Fa-1	Distribution of Delays	The system should not unfairly favor specific regions, connections, or groups of individuals. This means that when system disruptions cannot be avoided, they should be distributed fairly. Measures should be put in place to ensure that these constraints are observed.
Re-2	RUOM Favouritism	The system should not unfairly favor specific RUOMs. Re-scheduling in railway operations must impact the RUOMs fairly. Measures should be put in place to ensure that these constraints are observed.
O-1	Maintainability	Involves the ease with which the system can be maintained and updated. This includes the ability to diagnose and fix issues, update software, and adapt to changing operational requirements.
O-2	Environmental Sustainability	Addresses the system's impact on the environment. This includes considerations such as energy efficiency of the AI algorithms and the broader ecological footprint of the system's implementation and operation.

## 7 Common Terms and Definitions

Common Terms and Definitions	
Term	Definition
Railway Undertaking Operating Managers (RUOMs)	Company or organization that operates trains or provides rail transport services.
Traffic Management System (TMS)	It provides permanent control across the network, automatically sets routes for trains logs train movements, and detects and solves potential conflicts.

Co-learning	Co-learning indicates that human or AI in a team has the ability to interact, learn from/with, and grow with their collaborator. Co-learning aims to support two dynamic, growing entities to build mutual understanding, facilitate mutual benefit, and enable mutual growth over time. <i>Source: Huang, Y. C., Cheng, Y. T., Chen, L. L., Hsu, J. Y. J. (2019). Human-AI Co-learning for data-driven AI. arXiv preprint arXiv:1910.12544.</i>
Trains re-scheduling	Monitoring the movement of trains on a railway network and reacting to unexpected events, such as signal failures, track blockages, or weather events that disrupt operations, to other significant delays, and proactively to predicted deviations that affect planned operations. Re-scheduling measures include changing a train's speed, path, or platform for stopping.

## UC2.RAILWAY: AI-ASSISTED HUMAN RE-SCHEDULING IN RAILWAY OPERATIONS

### 1 Description of the use case

#### 1.1 Name of the use case

ID	Application Domain(s)	Name of Use Case
UC2.Railway	Railway network	AI-assisted human re-scheduling in railway operations

#### 1.2 Version management

Version Management			
Version No.	Date	Name of Author(s)	Changes
0.1	04.03.2024	Adrian Egli, Daniel Boos, Irene Sturm, Roman Ließner, Manuel Schneider, Julia Usher, Manuel Renold, Toni Wäfler, Samira Hamouche	Initial Version (import from UC2.Railway short)
0.2	15.04.2024	Anton Fuxjäger, Adrian Egli, Manuel Schneider, Julia Usher, Toni Wäfler, Roman Ließner, Cyrill Ziegler, Manuel Renold, Daniel Boos	Updated
0.3	16.04.2024	Ricardo Bessa	Revision
0.4	25.04.2024	Adrian Egli, Daniel Boos, Irene Sturm, Roman Ließner, Manuel Schneider	Final Revision
0.5	30.05.2024	Adrian Egli	Revision: Action space
1.0	08.07.2024	Ricardo Bessa	Final version

#### 1.3 Scope and objectives of use case

Scope and Objectives of Use Case	
Scope	Traffic density on the European rail networks is constantly increasing. This increases the complexity of rail traffic management in operations: timetables are

	<p>constructed to utilize the network’s capacity maximally. At the same time, new construction or maintenance of railway infrastructure must be planned and carried out efficiently. In railway operations, the already densely planned schedules are disturbed by unexpected events, such as delays, infrastructure defects, or short-term maintenance. The execution of the planned timetable can only be achieved by acting on these events by frequently adapting and re-scheduling the planned train runs. Already today, maintaining smoothly running operations requires that in operational centers, highly skilled personnel monitor the flow of traffic day and night and quickly make decisions about re-scheduling of trains.</p>
<p><i>Objective(s)</i></p>	<p>Aims to use AI-based methods to assist the human dispatcher in railway operations in re-scheduling train runs to fulfill all offered services and minimize delays for the customer (passenger).</p>
<p><i>Deployment model</i></p>	<p>Cloud services and on-premises.</p>



1.4 Narrative of use case

<i>Narrative of Use Case</i>
<p><b>Short description</b></p> <p>In railway operations, traffic on the network is planned to fulfill the intended service contracted with the Railway Undertaking Operating Managers (RUOM). In railway traffic operations, a pre-planned schedule is executed. Unexpected events, such as infrastructure malfunctions or delays, occur. In this use case, a disruption or deviation occurs, and a dispatcher needs to become aware of the situation, analyze it, and decide to fulfill the requested services as close as possible to the pre-planned schedule. In our case, the dispatcher should be supported by an AI-assisted system to choose some actions, e.g., changing the speed, order, or trains routes. The support system takes the state of all trains in the dispatcher’s control area as input and suggests options, i.e., sets of actions, to the dispatcher.</p>
<p><b>Complete description</b></p> <p>Train dispatching is responsible for managing the movement of trains across a complex rail network. Human dispatchers rely on a computerized dispatching system to plan and monitor train movements. However, unexpected disruptions, such as signal failures, track blockages, or weather events, can cause significant delays and disruptions to the train schedule. In the event of a disruption, dispatchers need to quickly make decisions to reschedule trains and minimize the impact on passengers and freight. This can be complex and time-consuming, especially considering the intricate network of tracks, train priorities, and passenger demand.</p> <p>In this use case, an AI-assistant system supports the human dispatcher. This system gets the real-time state of all the trains and tracks in the dispatcher’s control area and derives possible dispatching options in case of deviations from the pre-planned schedule due to disruptions or delays. The options are presented in near real-time to the dispatcher and consist of a set of actions the dispatcher can perform to bring the trains back or close to their pre-planned schedules.</p> <p>The following steps are performed in the use case:</p> <ol style="list-style-type: none"> <li>1. <b>Definition of system parameters:</b> Detailed parameters are set for the pre-planned schedule, including the prioritization of trains in case of disruptions, acceptable delay margins, and specific criteria for train prioritization (e.g., passenger load and destination importance). This step also includes the configuration of safety systems, network capacity limits, and any special operational requirements unique to certain routes or times.</li> <li>2. <b>Set up/configuration of human-AI teaming:</b> The human defines the boundary requirements, including the flexible allocation of decision-making authority between humans and machines.</li> <li>3. <b>Schedule execution:</b> The initial operational plan is put into action. This includes the deployment of trains according to the pre-planned schedule, monitoring of train movements, adherence to the sequence of commercial stops, and ensuring compliance with operational requirements like safety systems and traffic density management.</li> <li>4. <b>Monitoring:</b> At any time during operations, the human dispatcher can monitor the flow of traffic in the area of control. Visual displays of the traffic running through the network exist, and metrics are available. Information about the current intended plan is available.</li> <li>5. <b>Detection of deviation:</b> At any time in operations, the human-AI team detects an emerging deviation of the actual state of the system from the planned state. The re-scheduling process can be initiated by various triggers such as infrastructure changes (e.g., blocked tracks, malfunctioning switches), train delays, equipment malfunctions, or potential future issues. The system is designed to detect these deviations in real time and assess their impact on the overall schedule. The system also predicts issues that might become relevant in the future.</li> <li>6. <b>Action (re-scheduling):</b> Upon detecting a current or future deviation by the system or human, the system provides a detailed display of the issue, e.g., including its nature, location, and expected impact on the schedule. Either the human or the system starts with a suggestion, leading to two further paths of actions:</li> </ol>

- a. The system provides suggestions. The human provides feedback (e.g., context unknown to the system). AI adapts the solution based on the feedback. The human agent can choose to select one of the suggestions by the AI systems, initiate a new solution search, or choose their own course of action.
  - b. The human provides a suggestion. The AI system provides quantified feedback to the human suggestions, including own and adapted suggestions. Humans select one of the proposed solutions and initiate action. Alternatively, humans formulate a hypothesis, and the AI system provides evidence for and against these hypotheses.
7. **Execute solution:** The newly adapted schedule is implemented. The system continuously monitors for any further deviations and adjusts the schedule as needed to maintain operational efficiency and adherence to time constraints.
  8. **Human review and system adjustment:** A human supervisor reviews the system's performance, analyzing how effectively it responded to deviations and the impact on service delivery. Based on this review, adjustments are made to the system's parameters, such as altering the prioritization criteria, adjusting acceptable delay thresholds, or refining the algorithm for schedule recalculations. This step ensures continuous learning and improvement of the system based on operational experiences and organizational goals.
  9. **Co-learning:** AI agent learning loop using observations of the human decision-making process. The human learning process (e.g., to detect emerging deviations or to develop solutions) is explicitly supported by human-AI interaction.

**Stakeholders**

**Railway network operator:** Operator of the railway network in charge of maintaining traffic flow on the railway network to provide high quality-of-service to their direct customers (RUOMs) and the passengers.

**Network supervisor:** Human supervisor of the automated railway system (something like the former dispatcher who is not dispatching himself anymore but monitoring the system state),

**RUOM:** Railway Undertaking Operation Manager offering passenger and freight traffic services.

**Neighboring areas of control/operational centers.**

**Passenger:** The primary end-user of the railway services whose travel experience and satisfaction are directly impacted by the efficiency and punctuality of train operations.

**Government and society:** The quality of railway services is a concern of the government and society.

**Stakeholders' assets, values**

**Railway network operator:**

- Available capacity on the network: a low-quality re-scheduling functionality will consume more capacity on the network.
- Reputation: low performance of the AI system can lead to a bad reputation in terms of operational stability, punctuality, etc., which might cause customers to not rely on and to use less the services offered. This also concerns network operators, RUOM, and passengers.
- Legal and regulatory framework: Regulations with discrimination-free treatment of RUOMs.
- Unintended behavior of the AI system and actions by malicious actors can potentially compromise the safety of the train passengers, personnel on the train, and on and in proximity to the tracks, as well as infrastructure like tracks, power lines, tunnels, stations, etc.

**Human dispatcher:**

- Damage to the reputation as well as a potential general perception of an opaque AI system being in control of running trains can cause a decrease in the trustworthiness of the railway operator from a customer perspective, both for individual travelers and cargo transport.

The usefulness and understandability of the AI-system output to the dispatcher may influence the trustworthiness of the AI-system from the perspective of the dispatcher. Low trustworthiness might render the use of the AI system irrelevant as the dispatcher will not trust the options generated by the system, and the assumed benefit will not materialize.

***System's threats and vulnerabilities***

**Trust from human operators:** The operational performance of the AI assistant will not be close to 100% of problems solved, which may hinder the confidence and trust of the human operator in the AI recommendations. This could introduce a negative cognitive bias in humans.

**Progressive deviation of environment behavior:** Not only can the system conditions evolve but also the operational rules, the human operators' behavior, or other applicable regulation. This can progressively alter the efficiency of the AI assistant if it is not regularly "updated". The issue can be exacerbated by the fact that such changes happen very incrementally in time and are quite hard to detect at the early beginning, where only a few changes should be adopted.

**A mismatch between AI training and deployment:** Where significant differences exist between the digital environment used to train the AI model or the lack of information in historical data used to train the AI model can cause issues under real operating conditions. This could lead to low robustness and poor performance during execution, e.g., recommendations based on inaccurate assumptions about observability and controllability.

**Security:** The AI system introduces the risk of malicious actors disrupting operations either through the disabling or disruption of the AI system or by influencing system to produce output that causes delays, etc

### 1.5 Key performance indicators (KPI)

Name	Description	Reference to the mentioned use case objectives
Assistant relevance	Situation awareness of the human operator using the system It is based on an evaluation by the human operator of the relevance of action recommendations provided by the AI assistant and measured by the number of recommendations from the AI assistant effectively used by the human operator.	Linked to the capacity of the AI system to support the dispatcher in choosing some actions.
Human Information Processing	The volume of information that the human takes into account when making decisions with AI support (as compared to making decisions with no AI support).	Linked to the cognitive load of human dispatchers.
Punctuality	An aggregated measure of the delay in a scenario (defaults to be defined).	Linked to the objective of minimizing delays.
Response time	The time needed to produce a new schedule in case of a disturbance event.	Related to the objective of rapid re-scheduling.
Comprehensibility	It is defined as the ability to understand a decision logic within a model and, therefore, the ability to use this knowledge in practice (Futia and Vetrò, 2020). Futia, G. and Vetrò, A. (2020). On the Integration of Knowledge Graphs into Deep Learning Models for a More Comprehensible AI. <i>Information</i> , 11 (2), 122-132. Herm, L. V., Wanner, J., Seubert, F., & Janiesch, C. (2021). I Don't Get IT, but IT seems Valid! The Connection between Explainability and Comprehensibility in (X) AI Research. In ECIS.	Linked to interpretation of what has been learned and decision logic.
Acceptance	Acceptance of the system by a human user (e.g., Using the TAM model (technology acceptance model)).	Reflects the reliability and trust of the AI system.
Trust towards the AI-Tool	<i>“(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)”</i> (Cahour & Forzy, 2009, p. 1261). The human operators' trust towards the AI tool can be measured using the Scale for XAI (Hoffman et al., 2018) or similar.	Linked to the human operator's appropriate trust in the AI system as a necessary precondition of adequate use.
Human motivation	<i>“Intrinsic motivation is defined as the doing of 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”</i> (Ryan & Deci, 2000, p. 54).	This is linked to the necessary motivation of the human operator to use the AI for complete a task and reach corresponding objectives.

	<p>The human operators perceived internal work motivation can be measured by using the Job Diagnostic Survey (Hackman &amp; Oldham, 1974) or similar. The questionnaire needs to be adapted to the AI context (e.g., problem detection with AI-assistance).</p>	
Human control/autonomy over the process	<p>Autonomy is the degree to which the job provides substantial freedom, independence, and discretion to the employee in scheduling the work and in determining the procedures to be used in carrying it out” (Hackman &amp; Oldham, 1975, p. 162). It consists of three interrelated aspects centered on freedom in decision making, work methods and work scheduling (Morgeson &amp; Humphrey, 2006). Parker and Grote (2022) view job autonomy interchangeably with job control.</p> <p>The human operators perceived autonomy over the process can be measured by using the Work Design Questionnaire (Morgeson &amp; Humphrey, 2006) or similar. The questionnaire needs to be adapted to the AI context (e.g. problem detection with AI-assistance).</p>	<p>Linked to the perceived control of the human operator as a necessary prerequisite for taking responsibility for the efficiency and effectiveness of one’s own work.</p>
Human learning	<p>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, as well as the continuity over time (Alexander et al., 2009).</p> <p>The human operators perceived learning opportunities working with the AI-based system can be measured by using the task-based workplace learning scale (Nikolova et al., 2014) or similar. The questionnaire needs to be adapted to the AI context.</p>	<p>Linked to the objective of mutual co-learning to assist human operator to improve his/her performance.</p>
Decision support for the human operator	<p>Decision support tools should be aligned with the cognitive decision-making process that people use when making judgements and decisions in the real world and ensure that the human operator retains agency (Miller, 2023). AI decision support tools should therefore help people to remain actively involved in the decision-making process (e.g. by helping them critique their own ideas) (Miller, 2023).</p> <p>The decision support for the human operator can be measured based on the criteria for good decision support (Miller, 2023) or similar. The instrument needs to be further developed.</p>	<p>Linked to appropriateness of AI-based support of the human operator’s decision-making process.</p>
Ability to anticipate	<p><i>“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”</i> (Hollnagel, 2015, p. 4).</p> <p>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</p>	<p>Linked to AI-based enabling of human operator to minimize delays for the customers.</p>

	by an AI-based assistant can be measured by using the Rigor-Metric for Sensemaking (Zelik et al., 2010) or similar. The instrument needs to be further developed and adapted to the AI context.	
Situation awareness	<p>“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, p. 12).</p> <p>The human operator’s situation awareness can be measured by using the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1988) or similar.</p>	Linked to the AI-based assistance of the human operator for developing an appropriate situation awareness.

### 1.6 Features of use case

<b>Task(s)</b>	Planning, prediction, optimization, interactivity, recommendation
<b>Method(s)</b>	Reinforcement learning has been applied to this use case, but other AI approaches are possible.
<b>Platform</b>	<a href="#">Flatland</a> digital environment.

### 1.7 Standardization opportunities and requirements

<i>Classification Information</i>
<b><i>Relation to existing standards</i></b>
<p><b>ISO/IEC 23894:2023</b>, <i>Information technology — Artificial intelligence — Guidance on risk management</i>. Autonomous management and optimization of railway scheduling in real-time are high-stakes tasks, and therefore, risk management specifically related to AI is fundamental.</p> <p><b>ISO/IEC 38507:2022</b>, <i>Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations</i>. Autonomous AI requires an analysis of governance implications and also a redefinition of the organization structure.</p> <p><b>ISO/IEC 24029-2:2023</b>, <i>Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for using formal methods</i>. Since artificial neural networks can be a component of the autonomous AI system, formal methods to assess the robustness properties of neural networks are fundamental to certify and monitor autonomous systems.</p> <p>In railway transport, there are different levels of automation (Grade of Automation, GoA) defined in the <b>IEC 62267</b> Standard ("Railway applications - Automated urban guided transport (AUGT) - Safety requirements"). This standard covers high-level safety requirements applicable to automated urban guided transport systems, with driverless or unattended self-propelled trains, operating on an exclusive guideway.</p> <p><b>DIN EN 50126</b>, <i>Railway Applications – The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS)</i>. It considers the generic aspects of the RAMS life cycle and provides a description of a Safety Management Process. It provides guidelines for defining requirements, conducting analyses, and demonstrating the reliability, availability, maintainability, and safety aspects throughout the lifecycle of railway applications.</p> <p><b>DIN EN 50128</b>, <i>Railway applications – Communication, signaling and processing systems</i>. Outlines the procedural and technical criteria for crafting software intended for programmable electronic systems in railway control and protection applications.</p>
<b><i>Standardization requirements</i></b>
<p>Opportunities for standardization and deriving recommendations for critical operations management and support, especially regarding co-decision-making and human-computer interaction, as well as safety requirements. See also UC1.Railway.</p>

### 1.8 Societal concerns

<i>Societal concerns</i>
<b><i>Description</i></b>
<p><b>Privacy and data protection:</b> The use of AI in railway scheduling involves the collection and analysis of large volumes of data, including potentially sensitive information. There is a concern about how this data is stored, processed, and protected, especially in compliance with data protection regulations like GDPR. Ensuring the privacy and security of passenger and employee data is paramount.</p> <p><b>Transparency and accountability:</b> There is a societal demand for transparency in how AI systems make decisions, especially in critical infrastructure like railway systems. The public might be concerned about the lack of understanding of AI decision-making processes and the accountability mechanisms in place in case of failures or errors.</p>

<p><b>Employment and skill shift:</b> The automation of train scheduling might lead to concerns about job displacement and the need for reskilling of railway staff. While AI can optimize operations, it also changes the nature of work, requiring a shift in skills for human operators who now need to oversee and interact with advanced AI systems.</p> <p><b>Public trust and acceptance:</b> For the successful implementation of AI in public transportation, gaining and maintaining public trust is crucial. There may be apprehensions and resistance from the public regarding the shift to AI-driven systems, especially among those accustomed to traditional methods.</p> <p><b>Safety and security:</b> The use of AI-systems for critical operational scenarios raises concerns regarding the continued safety and security of these systems. Understanding failure modes, developing robust models, and ensuring resilience to adversarial attacks are among the many topics to be tackled.</p> <p><b>Inequality:</b> Such systems might introduce inequality in service quality for different geographic regions or categories of passengers due to the opacity of the system, bias and self-learning aspects.</p>
<p><b>Sustainable Development Goals (SDG) to be achieved</b></p>
<p>SDG9. Decent work and economic growth / SDG9. Industry, innovation and infrastructure / SDG11. Sustainable cities and communities / SDG13. Climate action</p>

## 2 Environment characteristics

<i>Data characteristics</i>	
<i>Observation space</i>	<p>Partially observable with limitations due to the unpredictable duration of delays and malfunctions.</p> <p>Data update is near real-time (rather seconds than hours).</p> <p>Domain: defined on a continuous space.</p> <p>Size: Depending on the type of observation considered local or global the total size can depend, but will generally be very large.</p> <p>Noise: The observation can be noisy due to the communication system and the various signaling devices (<i>signal box</i>).</p> <p>(In addition to more than 10,000 trains (per day), there are over 32,000 signals and over 14,000 switches in the Swiss rail network. All of this information must be taken into account and observed, thus the global observation is very large.)</p>
<i>Action space</i>	<p>The action space of the environment is mixed. Actions like which route to take on a switch are discrete as well as decisions like if a train should accelerate or decelerate. However, dependent on the algorithmic approach, the rate of acceleration, deceleration, the velocity to move forward and similar can be modelled both discrete and continuously.</p> <p>Also dependent on the algorithmic approach is the dimension of the action space. While the action space grows linearly with the number of trains for the algorithmic part, it grows exponentially if there is a central actor controlling all the trains. The action space of the human dispatcher is in any case exponentially growing with the number of trains.</p> <p>Further, the dimensionality of the action space depends on infrastructure and timetable elements like switches, signals and scheduled stops. Hereby, the impact on the dimensionality of the action space depends not only on the nature of the actor in the algorithmic part but also on the type of task, i.e. if the task is tackled episodically or sequentially on the algorithmic side. For the human dispatcher, the task is generally considered to be sequential, since an action is usually dependent on previous actions taken.</p> <p>Time horizon: for an action is typically from a few minutes to a couple of hours.</p> <p>The action space of the flatland environment is 5 (go left, go forward, go right, stop, none). However, each train run (agent) must perform one of these basic actions at each decision point (time step). This means that the total number of actions to be</p>



	selected is very large and stays in linear relation to the number of agents - i.e. in a problem-solving scenario with $n$ agents and $m$ time steps, the actions should be chosen in such a way that the combination of selected actions leads to the desired outcome or optimal solution. Each agent has a set of actions to choose from, from which they must select one at each time step. Therefore, the solution involves $n \times m \times a$ possible actions. (Up to 800 trains run simultaneously on the Swiss rail network. In many cases they interact directly or indirectly with each other.)
<i>Type of task</i>	The nature of the task depends on the algorithmic approach. While AI models can determine which action to take fully based on the current state without including information about past actions and would therefore be considered episodic, other approaches can, to a large degree, approach the problem-solving as a sequential task, for example, if planning is involved. The human dispatcher usually approaches the task sequentially.
<i>Sources of uncertainty</i>	Stochastic, with the following sources of uncertainty: <ol style="list-style-type: none"> <li>1) Weather conditions can impact, e.g. the friction of wheels on rails which leads to different acceleration and deceleration behavior.</li> <li>2) The travel demand influencing both the total load of a train and the delay to board other passengers.</li> <li>3) Disruptions: Train level – locomotives or other rolling stock issue that may arise and results into a delay; Infrastructure level – signal malfunctions or construction sites.</li> <li>4) Sensors and communication level – a failure may introduce noise and uncertainty in the observation of the environment.</li> </ol>
<i>Environment model availability</i>	A specific model of the environment is not available. Although a good approximation of it can be achieved as the basic laws of physics are defined and clear. However, a model of the environment will be simplified in general and subject to uncertainty (see above).
<i>Human-AI interaction</i>	Co-learning between the human and AI: The interaction between humans and AI is crucial in this specific use case. The use case allows for bidirectional communication in the decision-making problem, enabling humans to both use the system as a supporting tool for making decisions and to provide additional context and feedback to the AI to make the decision.

### 3 Technical details

#### 3.1 Actors

<i>Actor Name</i>	<i>Actor Description</i>
Dispatcher	The dispatcher is a human responsible for monitoring and analyzing railway traffic. The main role is to ensure the safe and efficient movement of trains by controlling the flow of traffic and making decisions based on real-time information. The dispatcher determines the order of trains and may deviate from planned routes when necessary to accommodate unexpected situations or optimize the overall operation. The decisions play a crucial role in maintaining the smooth functioning of the railway system.
Traffic control system	The traffic control system collects information such as traffic signals, train positions, and current train speeds and also provides a human-machine interface for controlling ongoing traffic. The system's goal is to manage the flow of traffic efficiently, centrally, and safely. This necessitates the comprehensive collection of available information to

	effectively support the decision-making process, which is primarily performed by human dispatchers. Consequently, the traffic control system is vital and should be implemented with a human-centered approach unless a fully automated solution is available.
Train run (Driver)	A train run refers to the operation of a train on a specific route or journey from one station to another. It encompasses the entire process of a train traveling along its designated path, including departure from the originating station, intermediate stops (if any), and arrival at the destination station. The current position and speed of the train are communicated to the traffic control system.

## 4 Step-by-step analysis of use case

### 4.1 Overview of scenarios

Scenario conditions					
No.	Scenario name	Scenario description	Triggering event	Pre-condition	Post-condition
1	Reactive Re-Scheduling	The reactive re-scheduling by the human-AI team once a deviation or disturbance has already occurred.	An emerging disruption or deviation occurring (e.g.. blocked track, malfunction train)	Intended service: a set of train runs with Start-and end location, a sequence of commercial stops, both with time information (Latest arrival, minimal dwell time, earliest departure). An initial (microscopic) operational plan that is executable and fulfils the intended services such as the arrival and departure times of trains at commercial stops.	System has produced a new operation plan that is executable in the simulation and leads to an “acceptable” state at the end of the scenario
2	Co-learning for reactive re-scheduling	The co-learning process initialized by the reactive re-scheduling by the human-AI team once a deviation or disturbance has already occurred.	Human and AI action and interaction during the re-scheduling process occurring after a disruption or deviation.	Initial human expertise and initial AI model required for corrective problem solving (e.g. solution generation).	Improved human expertise and/or improved AI model required for corrective problem solving. The improvement was the result of human-AI interaction explicitly supporting the human’s and/or the AI’s learning processes.
3	Proactive re-scheduling	Proactive re-scheduling by the human-AI team upon detection of weak signals.	Detection of precursors or weak signals indicating a probability of larger disruptions and deviation in the future	Intended service: a set of train runs with Start-and end location, a sequence of commercial stops, both with time information (Latest arrival, minimal dwell time, earliest departure). An initial (microscopic) operational plan that is executable and fulfils the intended services such as the arrival and departure times of trains at commercial stops.	System has produced a new operation plan that is executable in the simulation and leads to an “acceptable” state at the end of the scenario without the presence of any additional weak signals.
4	Co-learning for proactive re-scheduling	Co-learning process initialized by the proactive re-scheduling	Human and AI agent action and interaction during	Initial human expertise and initial AI model required for preventive problem solving (e.g.	Improved human expertise and/or improved AI model required for preventive problem solving.

		By the human-AI team	the detection and rescheduling phases.	problem detection, identification of leverage points).	The improvement was the result of human-AI interaction explicitly supporting the human's and/or the AI's learning processes.
--	--	----------------------	--	--	--

4.2 Steps of scenario

Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged	Requirement
1	Start	Definition of system parameters	Detailed parameters are set for the pre-planned schedule, including the prioritization of trains in case of disruptions, acceptable delay margins, and specific criteria for train prioritization (e.g., passenger load, destination importance). This step also includes the configuration of safety systems, network capacity limits, and any special operational requirements unique to certain routes or times.	Administrator	Network Operator	SYSPAR	
2	System params defined	Set up / configuration of human-AI teaming	The human defines the boundary requirements, including the flexible allocation of decision-making authority between human and machine.	Dispatcher	AI Assistant	CONFIG	
3	Teaming initialized	Schedule execution	The initial operational plan is put into action. This includes the deployment of trains according to the pre-planned schedule, monitoring of train movements, adherence to the sequence of commercial stops, and ensuring compliance with operational requirements like safety systems and traffic density management.	Dispatcher	TMS	EXECPLAN	

4	Information presented	Monitoring	At any time in operations the human dispatcher can monitor the flow of traffic in the area of control. There exist visual displays of the traffic running through the network and in addition metrics are available. Information about the current intended plan is available.	AI Assistant	Dispatcher	STATE	
5	Deviation detected	Detection of deviation	<p>At any time in operations an emerging deviation of the actual state of the system from the planned state is detected by the human-AI team. The re-scheduling process can be initiated by a variety of triggers such as infrastructure changes (e.g., blocked tracks, malfunctioning switches), train delays, equipment malfunctions or potential future issues. The system is designed to detect these deviations in real-time and assess their impact on the overall schedule. The system also predicts issues that might become relevant in the future.</p> <p>For scenarios 1 and 3, this step consists of <b>detecting</b> deviations (reactive) which have already occurred. In scenarios 2 and 4, the human-AI team <b>predict</b> (proactive) potential deviations. These detected / predicted deviations then trigger re-scheduling.</p>	AI Assistant/ Dispatcher	Dispatcher / AI Assistant	DEVINFO	

6	Suggestion provided	Re-scheduling suggestion	<p>Upon detecting a current or future deviation by the system or human, the system provides a detailed display of the issue, including its nature, location, and expected impact on the schedule. Either the human or the system starts with a suggestion, leading to two further paths of actions:</p> <p>The system provides suggestions. The human provides feedback (e.g., context that is not known to the system). AI adapts the solution based on the feedback. The human agent can choose to select one of the suggestions by the AI systems, initiate a new solution search, or choose their own course of action.</p> <p>The human provides a suggestion The AI system provides quantified feedback to the human suggestions, including own and adapted suggestions. Human selects one of the proposed solutions and initiate's action. Alternatively, the human formulates hypothesis, the AI system provides evidence for and against these hypothesis.</p>	AI Assistant/ Dispatcher	Dispatcher / AI Assistant	RESUG	
7	Suggestion received	Execute solution	<p>The newly adapted schedule is then put into operation. The system continuously monitors for any further deviations and adjusts the schedule as needed to maintain operational efficiency and adherence to time constraints.</p>	Dispatcher	TMS	RESCHED	

8	New schedule put into operation	Human review and system adjustment	A human supervisor reviews the performance of the system, analyzing how effectively it responded to deviations and the impact on service delivery. Based on this review, adjustments are made to the system's parameters, such as altering the prioritization criteria, adjusting acceptable delay thresholds, or refining the algorithm for schedule recalculations. This step ensures continuous learning and improvement of the system based on operational experiences and organizational goals.	AI Assistant	Dispatcher	REPORT	
9	Observation batch recorded / Training session	Co-learning	For scenarios 3 and 4, an additional co-learning loop occurs, consisting of a loop on the side of the AI agent and one on the side of the human agent. AI agent learning loop uses observations of the human decision-making process to improve its own decisions. Human learning process (e.g., to detect emerging deviations or to develop solutions) is explicitly supported by human-AI interaction.	TMS	AI Assistant / Dispatcher	OBS	



## 5 Information exchanged

<i>Information exchanged</i>		
<i>Information exchanged (ID)</i>	<i>Name of information</i>	<i>Description of information exchanged</i>
<a href="#">SYSPAR</a>	System Parameters	Series of parameters necessary to initialize the environment and providing all operative information to the agent(s).
<a href="#">CONFIG</a>	Configuration of human-AI teaming	Parameters defining the “work agreement” between the AI and human agent, for example the allocation of decision authority.
<a href="#">EXECPLAN</a>	Operational plan	Planned schedule to be executed, including information such as commercial stop sequence and operational requirements.
<a href="#">STATE</a>	State of the system	Detailed information on the current state of the system.
<a href="#">DEVINFO</a>	Devion information	Detailed information on the deviation, including its nature, location, and expected impact on the schedule.
<a href="#">RESUG</a>	Re-scheduling suggestions	Suggestion for rescheduling actions developed by the AI agent – e.g. list of actions to take in the next update cycles
<a href="#">RESCHED</a>	New operational plan	New schedule developed by the human-AI team.
<a href="#">REPORT</a>	Report of adjusted plan performance	Detailed performance report of system performance after executing the new operational plan, provided by the AI agent.
<a href="#">OBS</a>	Recorded observations	Series of rescheduling events and states including e.g. train run position, train run running state such as malfunctioning or good.

## 6 Requirements

<b>Requirements</b>		
<b>Categories ID</b>	<b>Category name for requirements</b>	<b>Category description</b>
Ro	Robustness	Encompasses both its technical robustness (ability of a system to maintain its level of performance under a variety of circumstances) as well as its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no unintentional harm can occur. <i>Source: EU-U.S. Terminology and Taxonomy for Artificial Intelligence. First Edition</i>
E	Efficiency	Ability of an AI system to achieve its goals or perform its tasks with optimal use of resources, including time, computational power, and data.
I	Interpretability	Make the behavior and predictions of AI systems understandable to humans, i.e., degree to which a human can understand the cause of a decision. <i>Source: Molnar, Christoph. Interpretable machine learning. Lulu.com, 2020.</i>

Re	Regulatory and legal	The AI system's capacity to meet its objectives while complying with relevant laws, regulations, and standards.
Fa	Fairness	Ensure the recommendations and predictions of the AI system are in line with the principles of fairness (i.e., fair distribution of the benefits and strain/harm)
O	Other	Other non-function requirements related to environmental concerns and maintenance
<b>Requirement R-ID</b>	<b>Requirement name</b>	<b>Requirement description</b>
Ro-1	Reasonable recommendations in new situations (not seen during model training)	Systems provides reasonable solutions for situations not seen during training.
Ro-2	Good performance in operating scenarios with high variability	System performs well in situations with many fast-changing elements.
Ro-3	Retrospective quality control	Quality of provided options can be assessed in retrospect.
E-1	Capacity to handle operating scenarios with high complexity	System derives options fast and with high quality in complex situations with many trains, switches and other elements involved.
E-2	Scalability	Concerns the system's ability to handle growth, such as increased train traffic or network expansion, without performance degradation. This ensures the system remains effective as the scale of railway operations increases.
E-3	Generalization to different scenarios	The system's ability to handle previously unseen scenarios and generalize to areas of observation and action space not visited during training (e.g., different speed profiles, rails configuration etc.)
I-1	Interpretability of suggestions	The process through which the AI system learns and operates, including how it generates suggestions, is transparent and understandable to the human dispatcher. Further, the decision making that leads to the suggestion as well as its limitations are explained to the human dispatcher.
Re-1	Compliance with legal standards and regulations	Adherence to data protection laws, safety regulations, cybersecurity, and ethical guidelines governing AI systems in public transportation and the EU AI Act.
Fa-1	Distribution of Delays	The system can be analysed to understand the distribution of delays according to certain fairness criteries (eg. region, RUOMs, groups, individuals) and allows to take measures to increase the fair distribution of delays.
Re-2	RUOM Favouritism	The system should not unfairly favour specific RUOMs. Re-scheduling in railway operations must impact the RUOMs according to official policy. Measures should be put in place to ensure that these constraints are observed.
O-1	Maintainability	Involves the ease with which the system can be maintained and updated. This includes the ability to diagnose and fix issues, update software, and adapt to changing operational requirements.

O-2	Environmental Sustainability	Addresses the system's impact on the environment. This includes considerations such as energy efficiency of the AI algorithms, and the broader ecological footprint of the system's implementation and operation.
-----	------------------------------	---

## 7 Common Terms and Definitions

Common Terms and Definitions	
Term	Definition
Railway Undertaking Operating Managers (RUOMs)	Company or organization that operates trains or provides rail transport services.
Traffic Management System (TMS)	Provides permanent control across the network, automatically sets routes for trains and logs train movements as well as detects and solves potential conflicts.
Co-learning	Co-learning indicate that human or AI in a team has the ability that can interact and learn from/with, and grow with their collaborator. The goal of co-learning is to support two dynamic growing entities to build mutual understanding, facilitate mutual benefit, and enable mutual growth over time. <i>Source: Huang, Y. C., Cheng, Y. T., Chen, L. L., Hsu, J. Y. J. (2019). Human-AI Co-learning for data-driven AI. arXiv preprint arXiv:1910.12544.</i>
Trains re-scheduling	Monitoring the movement of trains on a railway network and reacting to unexpected events, such as signal failures, track blockages, or weather events that disrupt operations, to other significant delays, and proactively to predicted deviations that affect planned operations. Re-scheduling measures include changing a train's speed, path, or platform for stopping.

## UC1.ATM: AIRSPACE SECTORISATION ASSISTANT

### 1 Description of the use case

#### 1.1 Name of the use case

ID	Application Domain(s)	Name of Use Case
UC1.ATM	Air Traffic Management	Airspace sectorisation assistant

#### 1.2 Version management

Version Management			
Version No.	Date	Name of Author(s)	Changes
0.1	04.12.2023	Clark Borst (TUD)	Initial document
0.2	15.01.2024	Clark Borst (TUD)	Major revision
0.3	03.02.2024	Ricardo Bessa	Revision
0.4	26.02.2024	Cristina Félix	Revision
1.0	15.04.2024	Cristina Félix	Final revision with new KPI's and ATM Workshop feedback update
1.1	12.05.2024	Clark Borst	Update scenario details with steps
1.2	14.06.2024	Anna Fedorova	Update
1.3	19.06.2024	Cristina Félix Joaquim Gerales Tiago Lima Reis	Final Revision
1.4	08.07.2024	Ricardo Bessa	Final version

#### 1.3 Scope and objectives of use case

Scope and Objectives of Use Case	
<b>Scope</b>	<p>Air traffic density in European airspaces is steadily increasing. At the same time, pressing economic and environmental concerns force a fundamental shift towards time- and trajectory-based air traffic operations. Taken together, increased traffic loads and operational complexities may eventually drive the workload peaks of the tactical air traffic controller (ATCO) beyond acceptable thresholds, threatening the overall safety of the Air Traffic Management (ATM) system and hindering a smooth transition toward a sustainable future of ATM.</p> <p>Solutions to manage the workload of ATCOs can already be applied in pre-tactical phases, for example, by splitting a large Flight Information Region (FIR) into several smaller airspace sectors that each are under the control of a single ATCO. Generally, pre-tactical ATM Sector Management ensures optimal sector configurations are always used to split traffic (and workload) over more ATCOs during tactical operations. Sectorisation is primarily meant to better handle daily traffic fluctuations, making optimal use of the personnel available.</p> <p>Today, sectorisation is the sole responsibility of the ATC supervisor, who exclusively decides <i>when</i> and <i>how</i> to split and merge sectors best, warranted by situational demands and available ATCO personnel. Only scattered information is available on different platforms to aid supervisors in this task. Still, there is currently no traffic</p>

	pre-analysis tool and/or integrated decision-support system to assist in or even fully automate the sectorisation process.
<b>Objective(s)</b>	The system's objective is to partially and 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.

### 1.4 Narrative of use case

<i>Narrative of Use Case</i>	
<b>Short description</b>	
<p>At ATC centers, such as Santa Maria Oceanic Area Control Centre (OACC), as part of NAV Portugal ANSP, a staff manager (i.e., ATC supervisor) exclusively decides <i>when</i> and <i>how</i> to split best and merge sectors, warranted by situational demands and available ATCO personnel. The degrees of freedom in sectorization involve considering horizontal (2D geometry) and/or vertical (altitude) constraints and can thus result in split sectors horizontally and/or vertically.</p> <p>Typically, under nominal conditions, the supervisor can install several pre-fab sectorization options. However, unexpected events, such as deteriorated weather conditions, flight emergencies (e.g., aircraft equipment failure), and unscheduled ATC personnel shortages (e.g., due to sickness) may require non-standard sectorisations to be installed.</p> <p>An AI-assistant, capable of operating under various levels of automation, will provide recommendations or even execute decisions on how to split the sector best horizontally, vertically, or both to balance ATCO workload while ensuring safety (i.e., adhere to horizontal and vertical separation criteria) and efficient traffic flows (i.e., reduce inefficiencies in flown track miles). The AI-assistant will also act in a bidirectional way by allowing the human operator to nudge the AI-generated recommendations in directions that seem more favorable.</p>	
<b>Complete description</b>	
<p><b>Description of the sectorization task:</b> Sectorisation involves retrieving and integrating several data information sources that are often gathered from different (digital) platforms, such as:</p> <ul style="list-style-type: none"> <li>• Expected traffic counts (available from EUROCONTROL CFMU)</li> <li>• Air-ground and coordination message count</li> <li>• Weather Information (METEO fore- and now casts)</li> <li>• Airspace Reservations (e.g., military airspace, temporary 'no-fly' zones)</li> <li>• Coordination Complexity (e.g., between area and arrival controllers)</li> <li>• Terminal Area Complexity (e.g., weather-related airport capacity limitations)</li> <li>• Equipment issues (e.g., communication issues between pilots and air traffic controllers)</li> <li>• ATCO staff schedules (depending on traffic demands)</li> </ul> <p>Based on the available ATCO personnel, including accounting for mandatory breaks after a 2.5-hour work cycle, the FIR is divided into several smaller airspace sectors, each under control by a single ATCO. How and when to best split and merge sectors horizontally and/or vertically depends on how well the traffic situation can be predicted over a specific time horizon. In general, the shorter the prediction horizon, the less uncertainty plays a role, but the more ad-hoc fluctuations in sectorisations can be expected with changing traffic loads. Therefore, a successful sectorization should be predictable and robust over a sufficiently long time horizon.</p> <p>At Santa Maria Oceanic Area Control Centre (OACC), as part of NAV Portugal ANSP, there are 3 pre-defined sectorization plans to be used by the supervisor under nominal operational conditions:</p> <ol style="list-style-type: none"> <li>1. <b>Unified Position.</b> Used in low traffic and/or complexity situations – one ATCO is responsible for working the full airspace.</li> <li>2. <b>VHF sector and non-VHF sector.</b> Used medium/high traffic and/or complex situations. This sectorization is used mostly when there is much terminal traffic or high volume inside the VHF coverage area. If the situation justifies, there is the possibility to vertically split the non-VHF</li> </ol>	

sector into several sectors to adjust the workload accordingly. The supervisor can also horizontally split the VHF sector into 3 different zones.

3. **North sector and South/Planning sector.** Used in medium/high traffic and/or complex situations. This sectorization is used mostly in low terminal traffic or low volume inside the VHF coverage area but with a high frequency of inbound coordination messages. If the situation justifies, there is the possibility of removing the VHF sector, creating 1 north sector, 1 south/planning sector, and 1 VHF coverage sector. In any of the examples above, the staff manager can split the South sector into several vertical sub-sectors.

There might exist more unexplored sectorization options, especially for novel/off-nominal operational conditions. In addition, the ATM community expects ATC staff shortages in the near future, requiring more flexibility in sector organizations. A hybrid AI system, based on supervised and unsupervised AI methods, could predict and provide sectorization solutions for nominal and off-nominal situations by learning from historical data and exploring new sector structures based on synthetic data generation.

**System description and role of the human operator:** The sectorization task is performed in a highly automated manner by an AI-based system. This system automatically observes the real-time data from all relevant ATM platforms, makes predictions on how and when to sectorise, and implements prediction results either as recommendations (to the human supervisor) or automatically installs the sectorization plan and bypasses the human. The AI system can be considered a new tool supervised and evaluated by a human expert. The AI system communicates its decisions on an auxiliary display that, for example, visualizes sector configurations on a map-like interface.

The role of the human operator (here, the ATC supervisor) is to evaluate the AI-based recommendations by requesting additional information and explanations, accepting or rejecting advisories, and nudging AI decisions in a different direction by manual interventions. All decisions and interactions will be logged, allowing the AI system to learn from human preferences continuously.

**Steps involved in the use case.** The following steps are performed in the ATM sectorization use case:

1. **Definition and identification of the critical system parameters.** Here, the relevant ATM system and contextual data needed for the sectorization task are gathered from (various) digital ATM platforms and integrated into a coherent, time-stamped “feature space” that drives sectorization predictions. Training and validation of the AI system are based on historical and synthetic/artificial data.
2. **Sectorisation implementation:** Based on predicted traffic, environment, and staffing conditions, a sectorization plan is predicted. The solution is presented to the human supervisor as a recommendation on an auxiliary interface. When the AI system acts at a lower level of automation, the human supervisor manually implements the sector plans. At higher levels of automation, the AI recommendations are executed based on “management by consent” (= AI implements only when the human accepts) or “management by exception” (= AI implements, unless the human vetoes). At the highest level of automation, the AI system is automatically implemented, and humans can only revise the system’s decisions afterward.
3. **Triggering sectorization revisions:** (Significant) changes in traffic loads, environment conditions, and staff availability can all trigger sectorization revisions. Parameters and thresholds warranting revisions will need to be defined and should be configurable for operational scenario generation.
4. **Human review and adjustment:** Depending on the level of automation set for the AI system, the role of the human supervisor ranges from manually implementing a sectorization plan to revising AI-implemented plans (see step 2). Humans can consult additional information and explanations underpinning AI’s decisions on demand, which is expected to foster trust in and acceptance of the AI system. As all human interactions are recorded, data will become available on what type of explanation is used most frequently and how certain explanations impact the acceptance of AI decisions. Such data can be used to improve the combined human-AI team performance.

#### Stakeholders

**ATC staff manager/supervisor:** The staff manager/supervisor located in the operational control room is responsible for the sectorization task.

<p><b>ANSPs responsible for the FIR:</b> e.g., NAV Portugal, the Portuguese Air Navigation Service Provider (ANSP), responsible for the Santa Maria Flight Information Region (FIR) and the Lisbon FIR.</p> <p><b>Other ANSPs:</b> Neighboring ANSPs, connected to the NAV FIRs (e.g., ONDA (Morocco) and ENAIRE (Spain)).</p> <p><b>Tactical Air Traffic Controller:</b> A single human ATCO responsible for maintaining safe, efficient and expeditious flows of air traffic within a single airspace sector.</p> <p><b>Airlines and pilots:</b> Airlines for adhering to planned operations; flight crew responsible for a safe and efficient execution of a planned flight.</p>
<p><b>Stakeholders' assets, values</b></p> <p><b>ATC staff manager / supervisor</b></p> <ul style="list-style-type: none"> <li>• Available personnel: Low-quality AI predictions may yield infeasible sectorization solutions (e.g., insufficient ATC personnel to handle all sectors)</li> <li>• Reputation: low performance of the AI system can lead to a bad reputation of the supervisor in devising workable and acceptable sectorisations (e.g., adhering to the mandatory ATCO breaks and preserving stability of a sectorization decision within a time window)</li> </ul> <p><b>ANSPs (incl. NAV and neighboring ANSPs)</b></p> <ul style="list-style-type: none"> <li>• Reputation: the ability to maintain efficient airspace usage and ability to coordinate traffic flows with neighboring FIRs</li> <li>• Safety: AI system recommendations should avoid creating traffic hotspots</li> </ul> <p><b>Tactical Air Traffic Controller (ATCO)</b></p> <ul style="list-style-type: none"> <li>• (Mental) workload and Situation awareness: AI-recommended sectorization should balance traffic loads in ways that adhere to acceptable workload limits and enable ATCOs to maintain situation awareness</li> </ul> <p><b>Airlines and pilots</b></p> <ul style="list-style-type: none"> <li>• Reputation: adhering to planned flights while reducing inefficiencies in flown track miles, possibly leading to delays</li> </ul>
<p><b>System's threats and vulnerabilities</b></p> <p><b>Accountability:</b> Who is responsible for the bad performance of the AI system</p> <p><b>Unexpected events:</b> Air traffic operations can be affected by events related to unexpected weather (e.g., local adverse weather cells, off-nominal wind conditions), flight emergencies (e.g., aircraft equipment failure), and unscheduled ATC personnel shortages (e.g., due to sickness). The scale of such events could lead to invalid or no solutions at all, for example, in the event of a volcano eruption or hurricane that requires closing off one entire airspace.</p> <p><b>Quality of data exchange infrastructure:</b> To ensure optimal decision-making, access to high-quality, real-time data will be required. Currently, information is scattered over various ATM systems, requiring a sufficiently robust IT infrastructure that can distribute data over the network to and from various Air Traffic Service (ATS) units. Delayed and uncertain information could negatively impact the quality of decisions.</p>

### 1.5 Key performance indicators (KPI)

<i>Name</i>	<i>Description</i>	<i>Reference to the mentioned use case objectives</i>
Acceptance score	Measure of acceptance degree of the generated AI solution for human operators	Reflects the acceptance choice in the AI's system decision. (0% - 100%). Measured directly from yes/no/revision input, translated into % across the operator's multiple interactions with AI-generated solutions

Agreement score	How much the supervisor agrees with the AI-generated sectorisation. Note: agreement and acceptance are not the same. One can accept a solution but not necessarily agree with it. A good system fosters a high-level agreement	This reflects the degree of agreement in the AI system. (Likert, 7-points scale)
Trust in AI solutions score	How much of the operator's confidence in the AI-generated solution, with and without the need for additional explanations.	This reflects trust in the AI system's decision. (Likert, 7-points scale)
Decision Support satisfaction	System effectiveness in supporting the efficient decision-making by airspace managers	Reflects the effectiveness of the AI system. (Likert, 7-points scale)
Efficiency score	How many times an AI-generated solution was revised. A good system would minimize the number of human interventions.	Reflects the efficiency of the combined human-AI team performance. (0% - 100%). Measured directly from user input (was the solution modified? Yes/no), translated into % across the operator's multiple interactions with AI-generated solutions
Significance of human revisions	The extent of human revisions compared to the AI decision. Here, small, localized revisions (e.g., merging two small adjacent sectors in the northeast corner of the FIR) would be rated differently from larger or multiple revisions across various areas in the FIR.	Reflects the AI system performance. (LOW, MED, HIGH interaction %). Measured directly from user input (of the modified solutions, how much interaction was measured? LOW number and extent of changes, MEDIUM number, and extent of changes HIGH number and extent of changes), translated into % across the operator's multiple interactions with AI-generated solutions
System Reliability	System trustworthiness - operation as expected under several conditions without major failures.	Reflects the efficiency of the combined human-AI team performance. (0%-100%). Measured directly from how many times the AI-generated solutions are sound or lead to failures
AI prediction robustness	How accurately and robustly does the AI system predict a certain sectorisation over a certain time horizon. Does re-evaluation of the sector structure in a shorter time horizon lead to different solutions? It is undesirable if small variations in capacity lead to significant differences in the sector structures/routings.	Reflects the efficiency of the combined human-AI team performance. Measured directly from the AI-generated solutions. How big a variation in capacity has to be to cause the AI to revise its previous solutions.
Prompt demand rate	Assess how many times the ATCO prompts additional explanations from the AI-generated solutions.	Reflects the AI system performance. (LOW, MED, HIGH interaction %) Measured directly from user input (how much interaction with explanations occurred and how the generated scenario is rated using the 'dynamic density index', measuring



		complexity), translated into % across the operator's multiple interactions with AI-generated solutions
AI co-learning capability	Does the ATCO feel that by the end of the trial runs, the AI has learned his preferences?	Links to the desired output of the AI system. (Likert, 7-points scale).
Human Response Time	Needed response time to react to AI advisory/information	(LOW, MED, HIGH response time %). Measured directly from user input (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), translated into % across the operator's multiple interactions with AI-generated solutions.

### 1.6 Features of use case

<b>Task(s)</b>	Planning, prediction, optimization, interactivity, recommendation.
<b>Method(s)</b>	Supervised Learning (e.g., ensemble decision trees) and possibly Reinforcement learning.
<b>Platform</b>	<a href="#">BlueSky</a> digital environment.

### 1.7 Standardization opportunities and requirements

<i>Classification Information</i>
<b><i>Relation to existing standards</i></b>
<p><b>ISO/IEC 23894:2023</b>, Information technology — Artificial intelligence — Guidance on risk management. Autonomous management and optimization of sectorisation in pre-tactical ATM operations are high-stake tasks, and therefore, risk management specifically related to AI is fundamental.</p> <p><b>ISO/IEC 38507:2022</b>, Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations. Autonomous AI requires an analysis of governance implications and also a redefinition of the organization structure.</p> <p><b>ISO/IEC 24029-2:2023</b>, Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for using formal methods. Since artificial neural networks can be a component of the autonomous AI system, formal methods to assess the robustness properties of neural networks are fundamental to certify and monitor autonomous systems.</p> <p><b>ICAO DOC 4444</b> – Standards and Recommended Practices in Air Traffic Management</p> <p><b>ERNIP Part 3</b> – EUROCONTROL Procedures for Airspace Management, Airspace Management Handbook for the Application of the Concept of the Flexible Use of Airspace. <a href="https://www.sesarju.eu/masterplan2020">https://www.sesarju.eu/masterplan2020</a> - European ATM Master Plan</p>
<b><i>Standardization requirements</i></b>
Establish a standard set of KPIs for measuring the performance of AI-based sectorisation systems and how the AI performance compares to heuristic methods in prediction and planning systems.

### 1.8 Societal concerns

<i>Societal concerns</i>
<b><i>Description</i></b>

<p><b>Increased air traffic density in Europe:</b> The challenge of maintaining safe and efficient air traffic management under increased traffic loads while adhering to the workload capacity limits of tactical ATCOs.</p> <p><b>Privacy and data protection:</b> The use of AI in ATM sectorisation involves the collection and analysis of large volumes of data, including potentially sensitive information. There is a concern about how data is stored, processed, and protected, especially in compliance with data protection regulations like GDPR.</p> <p><b>Transparency and accountability:</b> There is a societal demand for transparency in how AI systems make decisions, especially in high-stake transportation systems like ATM. The public might be concerned about the lack of understanding of AI decision-making processes and the accountability mechanisms in place in case of failures or errors.</p> <p><b>Employment and skill shift:</b> The full automation of the sectorisation task might lead to concerns about job displacement and the need for reskilling of ATC staff. While AI can optimize operations, it also changes the nature of work, requiring a shift in skills for human operators who now need to oversee and interact with advanced AI systems.</p> <p><b>Public trust and acceptance:</b> For the successful implementation of AI in air transportation, gaining and maintaining public trust is crucial. There may be apprehensions and resistance from the public regarding the shift to AI-driven systems, especially among those accustomed to traditional methods.</p>
<p><b>Sustainable Development Goals (SGD) to be achieved</b></p>
<p>SGD9. Industry, innovation and infrastructure / SGD11. Sustainable cities and communities / SGD13. Climate action</p>

## 2 Environment characteristics

<i>Data characteristics</i>	
<b>Observation space</b>	<p>Partially observable.</p> <p>Data updates are near real-time with a certain look-ahead time (minutes up to hours).</p> <p>Domain: defined on a continuous space.</p> <p>Size: &gt; 2000 flights per day, with &gt; 10 observable states per flight, &gt; 8 sectors with &gt; 20 coordination points (entry and exit points) per sector. .</p> <p>Noise: The observation of flight and sector data can be noisy due to unsynchronized update frequencies and data quality of various data platforms (e.g., meteo updates).</p>
<b>Action space</b>	<p>Mixed action space: sectorisation decisions are discrete (e.g., ‘split’ and ‘merge’), but sector geometry can vary on a continuous space depending on the algorithmic approach.</p> <p>Size: The action space of the human ATC staff manager is limited to the number of sectors to choose from and depends on ATCO staff availability, the number of flights, and the weather conditions (determining geographic restrictions)</p> <p>Time horizon: sectorisation actions range typically from a few minutes to a couple of hours (= pre-tactical operations)</p>
<b>Type of task</b>	<p>Human staff managers and AI assistants act in a sequential environment: the previous decisions can affect all future decisions. The next action of these agents depends on what action they have taken previously and what action they are supposed to take in the future.</p>
<b>Sources of uncertainty</b>	<p>Stochastic (weather forecasts, variability in traffic load, unpredicted ATCO staff shortage.)</p>
<b>Environment</b>	<p>Yes (aircraft performance models, ISA standard atmosphere)</p>

<i>model availability</i>	
<i>Human-AI interaction</i>	Co-learning between the human and AI: AI assistant proposes a sectorization plan, human evaluates plan, and human accepts or revises the plan (= feedback to AI assistant).

### 3 Technical details

#### 3.1 Actors

<i>Actor Name</i>	<i>Actor Description</i>
Staff supervisor	The human staff supervisor is responsible for implementing a sectorisation plan on a pre-tactical time scale. The staff supervisor needs to evaluate the outputs of an AI assistant that aims to support the staff manager in generating sectorisation suggestions.
AI assistant	The AI assistant provides sectorisation plan suggestions to the staff supervisor. It takes predicted information about the environment from various systems (e.g., weather forecasts from METEO services, traffic loads from Central Flow Management Unit, ATCO staff schedule, etc.) and historical data to aid the human staff manager. In the training phase, it can act on the environment to evaluate its recommendations. In the evaluation/testing phase, the actions on the environment should be performed by the human only.
Environment	The staff manager interacts with the BlueSky digital environment and with the AI assistant through a secondary interface. The AI assistant can also portray its sectorisation recommendations directly in the BlueSky environment (top-down Earth map).

## 4 Step-by-step analysis of use case

### 4.1 Overview of scenarios

Scenario conditions					
No.	Scenario name	Scenario description	Triggering event	Pre-condition	Post-condition
1	Nominal operational conditions	The AI sectorisation system responds to predicted traffic fluctuations under nominal operational conditions. Variations in traffic loads over a typical day (24 hours) will be used as inputs.	Fluctuations in traffic load over 24 hours, including periods of inbound and outbound of Santa Maria FIR.	Nominal ATCO staffing capacity	The system proposes and/or executes acceptable sectorisation results and presents results on an auxiliary interface for the human supervisor to evaluate.
2	Environment perturbations	This scenario deals with sudden changes in airspace availability due to adverse weather conditions of different magnitudes/scales, impacting sectorisation results.	Over a 24-hour period, various durations and scales of weather-related perturbations (e.g., off-nominal wind conditions due to storms) may require off-standard sectorisations.	Nominal ATCO staffing capacity	The system proposes and/or executes off-standard sectorisation results and presents results on an auxiliary interface for the human supervisor to evaluate.
3	ATCO staff shortage	This scenario deals with off-nominal ATCO staffing capacities, impacting sectorisation results.	Over a 24-hour period, various perturbations in ATCO staffing capacities (e.g., due to sickness) will require off-standard yet acceptable sectorisations. These events may be used in conjunction with environmental perturbations, simulating edge-case situations.	Off-nominal ATCO staffing capacity	The system proposes and/or executes off-standard sectorisation results and presents results on an auxiliary interface for the human supervisor to evaluate.

### 4.2 Steps for all scenarios

For each scenario the number of steps are the same and in-line with current practices in sectorisation on medium- to long-term time scales.

Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged
1	Start	The staff manager prepares his/her shift	The staff manager looks at estimated traffic counts and operational conditions and, using his experience, decides on 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.	Staff manager	AI assistant	SET
2	Initialise sector plan	AI assistant generates an initial sector plan	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, a timeline showing ATCO staff occupancy per sector, and a 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 conditions (e.g., wind) is also displayed and responsive to the time slider.	AI assistant	Staff manager	SPLAN
3	Plan evaluation	The staff manager evaluates the sector plan	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.	AI assistant	Staff manager	STATE
4	Human interacts	The staff manager interacts with the sector plan	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).	Staff manager	AI assistant	DEC

5	Re-schedule	Trigger an alert to re-schedule	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.	AI assistant	Staff manager	AL
---	-------------	---------------------------------	---	--------------	---------------	----

## 5 Information exchanged

<b>Information exchanged (ID)</b>	<b>Name of information</b>	<b>Description of information exchanged</b>
SET	Inputs and settings for AI assistant	Staff manager sets maximum time horizon and ATCO staff availability for the AI assistant
SPLAN	Sector plan	AI assistant suggestions for sectorisation.
STATE	Predicted system state	Predicted system state over a certain time period, including traffic load, weather conditions, ATCO shifts, sector topology, probability, and robustness score.
DEC	Human decision/interaction with the AI assistant operator	Staff manager's choice in terms of accepting, nudging, and revising.
AL	AI assistant alert	AI assistant issuing an alert, signaling to the staff supervisor that data used for predictions have changed significantly, warranting re-scheduling.

## 6 Requirements

<b>Requirements</b>		
<b>Categories ID</b>	<b>Category name for requirements</b>	<b>Category description</b>
Ro	Robustness	It encompasses both its technical robustness (the ability of a system to maintain its level of performance under a variety of circumstances) as well as its robustness from a social perspective (ensuring that the AI system duly considers the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no unintentional harm can occur. <i>Source: EU-U.S. Terminology and Taxonomy for Artificial Intelligence. First Edition</i>
E	Efficiency	The ability of an AI system to achieve its goals or perform its tasks with optimal use of resources, including time, computational power, and data.
I	Interpretability	Make the behavior and predictions of AI systems understandable to humans, i.e., the degree to which a human can understand the cause of a decision. <i>Source: Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.</i>
Re	Regulatory and legal	The AI system's capacity to meet its objectives while complying with relevant laws, regulations, and ethical standards.
O	Other	Other non-function requirements related to environmental concerns and maintenance
<b>Requirement R-ID</b>	<b>Requirement name</b>	<b>Requirement description</b>
Ro-1	System resilience to unexpected events	The AI system should work correctly under a variety of conditions and withstand operational disruptions. This includes resilience to

		unexpected events like adverse weather and sudden changes in the ATCO staff availability.
Ro-2	Cyber and data security	Focuses on protecting the system against unauthorized access, cyber threats, and data breaches. This ensures the integrity and confidentiality of sensitive operational data and safeguards the system from malicious attacks.
Ro-3	System's reliable operation and decisions	Shall show the capacity to perform its required functions under stated conditions for a specified period. This includes maintaining consistent performance and minimizing system failures or errors.
E-1	Capability to optimize resources and operations	The system shall maximize airspace and ATCO staffing utilization.
E-2	Scalability	Concerns the system's ability to handle growth in traffic loads, such as increased air traffic or airspace expansion, without performance degradation. This ensures the system remains effective as the scale of ATM operations increases.
I-1	Provide clear, understandable explanations for its decisions	It is crucial for human operators to validate and trust the AI's decisions, especially in complex sectorisation scenarios.
I-2	Usability of the system from the human and other stakeholders' perspective	It should include intuitive interfaces, ease of use, and effective communication of information.
Re-1	Compliance with legal standards and regulations	Adherence to data protection laws, safety regulations, and ethical guidelines governing AI systems in public transportation and the EU AI Act.
O-1	Maintainability	Involves the ease with which the system can be maintained and updated. This includes the ability to diagnose and fix issues, update software, and adapt to changing operational requirements.
O-2	Environmental Sustainability	Addresses the system's impact on the environment. This includes considerations such as energy efficiency of the AI algorithms and the broader ecological footprint of the system's implementation and operation.

## 7 Common Terms and Definitions

Common Terms and Definitions	
Term	Definition
Air Traffic Controller (ATCO)	Human operator, responsible for directing air traffic through a volume of airspace in a safe (i.e., maintaining separation standards) and efficient manner (i.e., expediting the flow of traffic, reducing delays, and avoiding inefficiencies in flow track miles).
Air Navigation Service Provider (ANSP)	Organization that provides the service of managing the aircraft in flight or in the maneuvering area of an airport and which is the legitimate holder of that responsibility. In this use case, NAV Portugal is the considered ANSP.
Flight Information Region (FIR)	A three-dimensional area in which aircraft are usually under the control of a single authority (ANSP). Sometimes, one or more FIRs have a combined upper area control, and/or FIRs are split vertically into lower and upper sections.



Airspace sector	A three-dimensional geographical area within an FIR is under control by a single ATCO or multiple ATCOs (e.g., planner and executive controller). Commonly, a FIR is divided into multiple sectors.
-----------------	---

## UC2.ATM: FLOW & AIRSPACE MANAGEMENT ASSISTANT

### 1 Description of the use case

#### 1.1 Name of the use case

ID	Application Domain(s)	Name of Use Case
UC2.ATM	Air Traffic Management	Flow & Airspace management assistant

#### 1.2 Version management

Version Management			
Version No.	Date	Name of Author(s)	Changes
0.1	15.01.2024	Clark Borst (TUD)	Initial document
0.2	19.01.2024	Joaquim Geraldés (NAVP) Cristina Félix (NAVP) Hélio Sales (NAVP)	Major revision
0.3	03.02.2024	Ricardo Bessa	Revision
0.4	05.02.2024	Joaquim Geraldés (NAVP) Cristina Félix (NAVP) Hélio Sales (NAVP)	Second major revision
0.5	13.02.2024	Giulia Leto (TUD) Clark Borst (TUD)	Revision and polishing
0.5.1	26.02.2024	Cristina Félix	Minor editorial change
1.0	15.04.2024	Cristina Félix	Final revision with new KPI's and ATM workshop feedback update
1.1	18.04.2024	Giulia Leto	Scenario updates with ATM workshop feedback
1.2	13.05.2024	Clark Borst	Update scenario details with steps
1.3	14.06.2024	Clark Borst	Update
1.4	19.06.2024	Cristina Félix Joaquim Geraldés Hélio Sales	Final Revision
1.5	08.07.2024	Ricardo Bessa	Final version

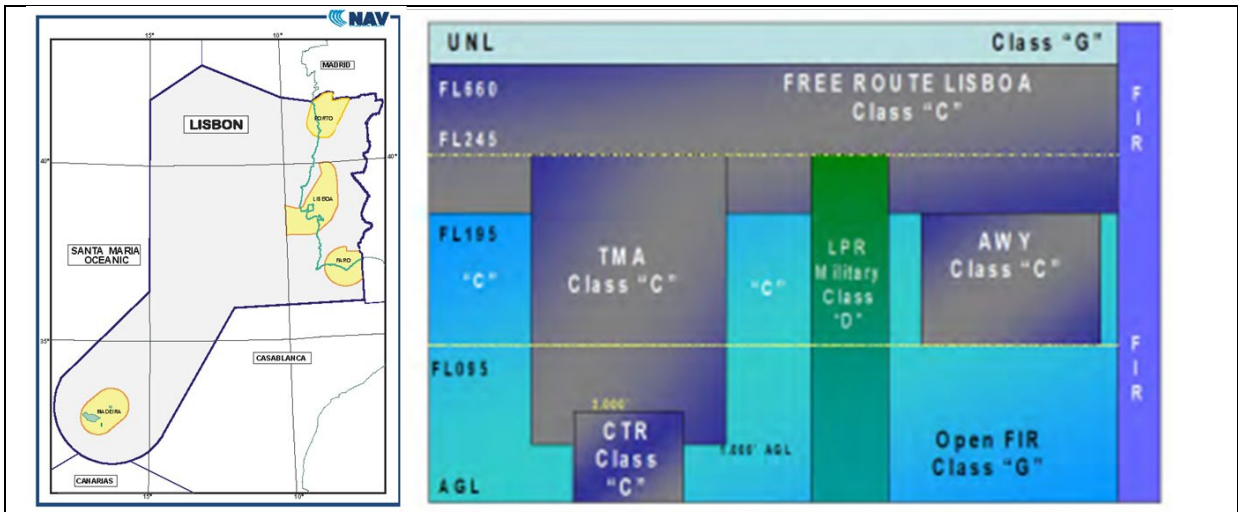
#### 1.3 Scope and objectives of use case

Scope and Objectives of Use Case	
Scope	<p>Air traffic density in European airspaces is steadily increasing. At the same time, pressing economic and environmental concerns force a fundamental shift towards time- and trajectory-based air traffic operations. Taken together, increased traffic loads and operational complexities may eventually drive the workload peaks of the tactical air traffic controller (ATCO) beyond acceptable thresholds, threatening the overall safety of the ATM system and hindering a smooth transition towards a sustainable future of ATM.</p> <p>For instance, in the Lisbon Flight Information Region (FIR), serviced by NAV Portugal, operational complexities arise from the activation of military areas, which</p>

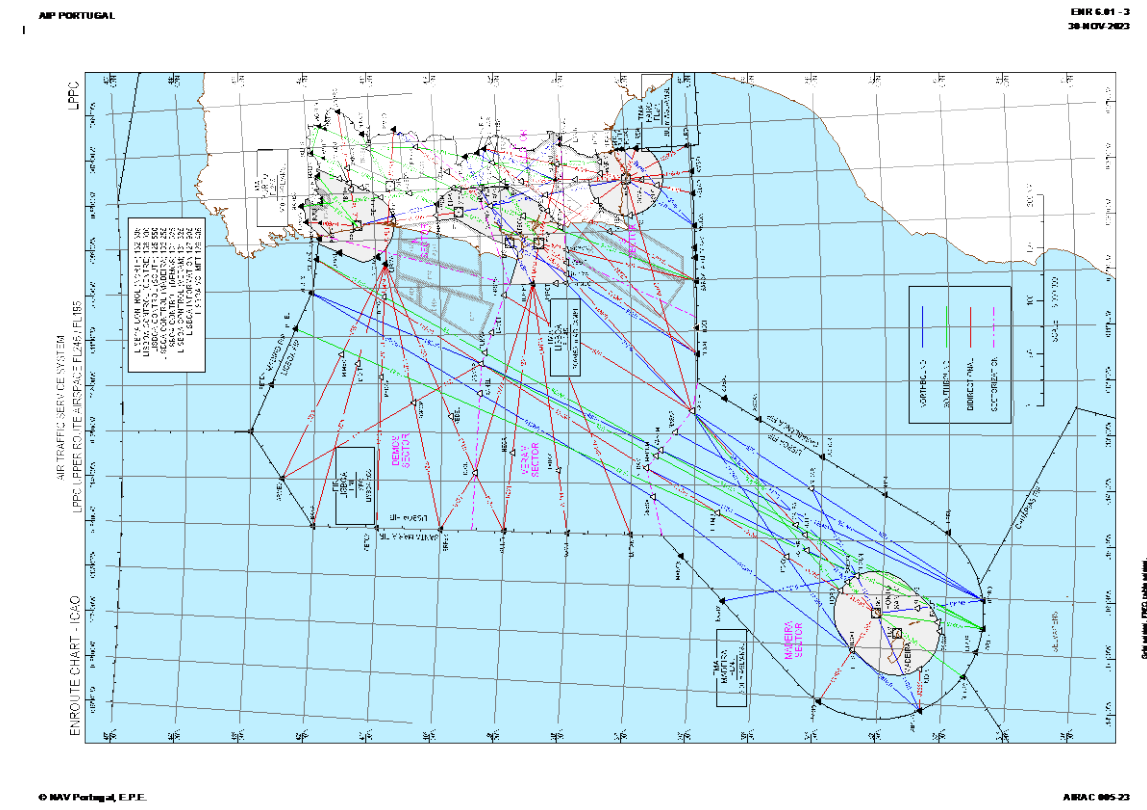
	<p>can significantly restrict the usage of the upper airspace for General Air Traffic (GAT), requiring traffic to deviate horizontally, especially when in combination with unexpected events (e.g. deteriorated weather conditions, flight emergencies). Routing of flight around military areas is proposed and implemented in pre-tactical phases. As of today, there is no pre-analysis tool and/or integrated decision-support system for assisting in, or even fully automating, the structuring of sectors with trajectory-efficient (e.g., flight time and fuel burn) routes and sectorisations to keep the workload of the tactical ATCOs within acceptable thresholds, i.e. without exceeding sector capacity limits.</p>
<p><b>Objective(s)</b></p>	<p>The system's objective is related to the flight execution phase when a military area is activated and the ATC has to issue deviations to avoid the activated area. The goal is to provide advice to ATCO about deviations with better sector capacity adherence and performance measured by an indicator of environmental area - <i>en-route flight inefficiency of the actual trajectory</i> (KEA). The use case will consider, as well, the need to review the sectorisation plan due to the military areas activation and required trajectory efficient deviations.</p>

**1.4 Narrative of use case**

<i>Narrative of Use Case</i>
<p><b>Short description</b></p>
<p>The Lisbon FIR includes an upper airspace area, four lower-airspace Terminal Maneuvering Areas (TMAs) and several military permanent and temporarily restricted areas. Because the upper Lisbon airspace is a so-called Free Route Airspace (FRA), flights can take any preferred route from entry to exit points, but preferably the most efficient (short) route.</p> <p>The activation/deactivation of military airspace in the Lisbon FRA can induce deviations from the flight plan routes. In this sense, to optimize the lateral deviation of the flights due to avoidance of an eventual temporary military activated area, the AI assistant will analyze and suggest a decision in sectorisation and routing of the main flows in Lisbon FIR (e.g., flight from London to Lisbon via either North or East entry coordination points of the Lisbon FIR).</p> <p>Human operators, more specifically the ATC and FMP supervisors, will be supported by an AI-assistant in how to best configure airspace sectors and optimize the routes for traffic flows at the enroute sectors of the Lisbon FIR in order to balance achievement of a better KEA (<i>Key performance Environment indicator based on Actual trajectory</i>, measuring the average en-route additional distance with respect to the great circle distance) and adherence to sector capacity limitations. The AI assistant will also act in a bidirectional way by allowing the human operator to nudge the AI-generated recommendations in more favorable/acceptable directions. The airspace sectorisation and flow structures, as devised by the AI and nudged by the operators in the pre-tactical phase, will be used by Tactical Air Traffic Controllers to manage traffic around the military activated areas.</p>
<p><b>Complete description</b></p>
<p><b>Description of the current Lisbon FIR situation:</b> The Lisbon FIR includes four TMA's (marked in yellow in the figure below). Within the Lisbon FIR, the airspace is classified "C", "D", and "G", with the airspace classification "D" being associated with military restricted areas. Under the Flexible Use of space (FUA) concept, the military-restricted areas may be released for management by the ANSP in order to allow for General Air Traffic (GAT) operations. When the military areas are released to the ANSP, the airspace classification of the delegated areas changes from "D" to "C".</p>



Above FL 245, the concept of Free Route Airspace in the Lisbon FIR (FRAL) is implemented since May 2009. Under the FRAL concept, all upper airspace of the FIR is available by default for civil aircraft planning purposes. Within the upper airspace, the activation/deactivation of military areas (highlighted with grey contours in the figure below) and its impact on civil planned flights is handled in the pre-tactical time horizon, as the activation of military areas can be planned from several weeks to one day in advance. Transitions from the upper Lisbon airspace to the TMAs in the lower Lisbon airspaces occur at fixed coordination points.



Currently, en-route flight inefficiency of the flown trajectories is monitored and targeted through a *Horizontal En-route Flight Efficiency KPI*, the *Key performance Environment indicator based on Actual trajectory (KEA)*. Routings deviating from those in nominal conditions, caused by military activations, changes in weather conditions or deviating airline decisions may lead to worse KEA values. As the Lisbon FIR above FL 245 is free of pre-defined routes, flexibility for routing outside of the restricted areas is available to account for major

deviations of the KEA. However, re-routing too many flights through the same airspace may exceed the sector capacity limit, requiring vertical and/or horizontal splits (i.e., sectorisations) to balance ATCO workload.

Therefore, given certain environmental and operational conditions, FRA structures and routings might exist that balance *flexibility* against *predictability* targets in optimized ways. Here, “optimized” is defined in terms of flight trajectory efficiency (e.g., flight time and fuel burn) and reduced operational complexity (e.g., crossing and merging points) that impact ATCO workload, leading in the worst case to exceed the sector capacity limits. A hybrid AI system, based on supervised and unsupervised AI methods, could analyse and provide routing and airspace configuration solutions for various operational scenarios in which the Lisbon FRA is restricted (due to activated military areas, weather conditions, etc.), predicting the KEA penalty and suggesting new routings and sectorisations that minimize the KEA while respecting sector capacity limits. Training scenarios can be selected from historical data and, for highly perturbed scenarios, can be based on synthetic data generation.

**System description and role of the human operator:** The airspace design for capacity and flow management for operational scenarios in which the Lisbon FRA is restricted is performed in a highly automated manner by an AI-based system. This system automatically observes data from all relevant ATM platforms and makes predictions on how to organize the airspace in terms of routings and sectorisation, and implements results as recommendations to the human operator (e.g., ATC and FMP supervisors).

The AI system can be considered as a new tool that is supervised and evaluated by a human expert. The AI system communicates its decisions on an auxiliary display that, for example, visualizes airspace configurations on a map-like interface.

The role of the human operator (here, the ATC and FMP supervisors) is to evaluate the AI-based recommendations by requesting additional information and explanations, accept or reject advisories, and nudge AI decisions in a different direction by manual interventions. All decisions and interactions will be logged, allowing the AI system to continuously learn from human preferences.

**Steps involved in the use case.** The following steps are performed in the ATM Flow & Airspace management use case:

5. **Definition and identification of the critical system parameters.** Here, the relevant ATM system and contextual data needed for the airspace structuring (i.e., routing and sectorisation) task are gathered from (various) digital ATM platforms and integrated into a coherent, time-stamped “feature space” that drives airspace structuring predictions. Training and validation of the AI system are based on historical and synthetic/artificial data.
6. **Airspace structuring implementation:** Based on predicted traffic, airspace military activations, environment, and staffing conditions, a minimum KEA routing plan and consequential sectorisation plan are predicted. The solution is presented to the human supervisor as a recommendation on an auxiliary interface. When the AI system acts at a lower level of automation, the human supervisor manually implements the routes and sector plans. At higher levels of automation, the AI recommendations are executed based on “management by consent” (= AI implements only when the human accepts) or “management by exception” (= AI implements unless the human vetoes). At the highest level of automation, the AI system is automatically implemented, and humans can only revise the system’s decisions afterward.
7. **Triggering airspace structuring revisions:** (Significant) changes, namely on military airspace activations & deactivations, as well as traffic loads, environment conditions, and staff availability, can all trigger routing and sectorisation revisions. Parameters and thresholds warranting revisions will need to be defined and should be configurable for operational scenario generation.
8. **Tactical deviations implementation:** Based on the operational conditions that lead to steps 2&3 above, the Tactical Air Traffic Controller will reroute the traffic around the military-activated areas to balance the better KEA and sector capacity adherence.
9. **Human review and adjustment:** Depending on the level of automation set for the AI system, the role of the human operator ranges from manually implementing a routing and sectorisation plan to revising AI-implemented plans (see step 2). Humans can consult additional information

and explanations underpinning AI's decisions on demand, which is expected to foster trust in and acceptance of the AI system. As all human interactions will be recorded, data will become available for the type of explanation used most frequently and how certain explanations impact the acceptance of AI decisions. Such data can be used to improve the combined human-AI team performance.

**Stakeholders**

**ATC supervisor**

The air traffic control supervisor, who is located in the operational control room, is responsible for the airspace-structuring task.

**FMP supervisor**

Local Flow Management Position supervisor is responsible for sector capacity management.

**ANSPs responsible for the FIR**

e.g., NAV Portugal, the Portuguese Air Navigation Service Provider (ANSP), responsible for the Santa Maria Flight Information Region (FIR) and the Lisbon FIR.

**Other ANSPs**

Neighboring ANSPs are connected to the NAV FIRs (e.g., ONDA (Morocco) and ENAIRE (Spain)).

**Tactical Air Traffic Controller**

A single human ATCO is responsible for maintaining safe, efficient, and expeditious flows of air traffic within a single airspace sector.

**National Air Force**

Example: the aerial military force of Portugal (Força Aérea Portuguesa (FAP)), responsible for the Air Search and Rescue Service, air policing service and Flight Information Service (FIS).

**Airlines and pilots**

Airlines for adhering to planned operations; flight crew responsible for the safe and efficient execution of a planned flight.

**Society and the general public**

Operational efficiency and safety pay dividends in terms of fuel burn, CO2 emissions, and punctuality.

**Stakeholders' assets, values**

**ATC or FMP supervisor**

- Available personnel: low-quality AI predictions may yield infeasible airspace structuring solutions (e.g., insufficient ATC personnel to handle all sectors).
- Tactical activations with short notice may affect the scenery (e.g., route efficiency decreases due to flight deviations, and the capacity of the sectors dedicated to GAT exceeded).

**ANSPs (incl. NAV and neighboring ANSPs)**

- Reputation: the ability to maintain efficient airspace usage and ability to coordinate traffic flows with neighboring FIRs.
- Safety: AI system recommendations should avoid creating traffic hotspots.

**Tactical Air Traffic Controller (ATCO)**

- (Mental) workload and Situation awareness: AI-recommended airspace structuring (routings of flights and sectorisation) should balance traffic loads in ways that adhere to acceptable workload limits and enable ATCOs to maintain situation awareness.

**Airlines and pilots**

<ul style="list-style-type: none"> <li>• Reputation: adhering to planned flights while reducing inefficiencies in flown track miles, possibly leading to delays.</li> </ul>
<p><b>System's threats and vulnerabilities</b></p> <p><b>Unexpected events:</b> Air traffic operations can be affected by events related to unexpected weather (e.g., local adverse weather cells, off-nominal wind conditions), flight emergencies (e.g., aircraft equipment failure), and unscheduled ATC personnel shortages (e.g., due to sickness). The scale of such events could lead to invalid or no solutions at all, for example, in the event of a volcano eruption or hurricanes that require closing off an entire airspace.</p> <p><b>Quality of data exchange infrastructure:</b> To ensure optimal decision-making, access to high-quality, real-time data will be required. Currently, information is scattered over various ATM systems, requiring a sufficiently robust IT infrastructure that can distribute data over the network to and from various Air Traffic Service (ATS) units. Delayed and uncertain information could negatively impact the quality of decisions.</p>

### 1.5 Key performance indicators (KPI)

<i>Name</i>	<i>Description</i>	<i>Reference to the mentioned use case objectives</i>
Acceptance score	Measure of acceptance degree of the generated AI solution for human operators	Reflects the acceptance choice of the AI's system decision. (0% - 100%). Measured directly from yes/no/revision input, translated into % across the operator's multiple interactions with AI-generated solutions.
Agreement score	How much the supervisor agrees with the AI-generated sectorisation. Note: agreement and acceptance are not the same. One can accept a solution but not necessarily agree with it. A good system fosters a high-level agreement	This reflects the degree of agreement on the AI system proposal. (Likert, 7-points scale)
Trust in AI solutions score	How much of the operator's confidence in the AI-generated solution, with and without the need for additional explanations.	This reflects trust in the AI system's decision. (Likert, 7-points scale)
Decision Support satisfaction	System effectiveness in supporting the efficient decision-making by airspace managers	Reflects the effectiveness of the AI system. (Likert, 7-points scale)
Efficiency score	How many times an AI-generated solution was revised. A good system would minimize the number of human interventions.	Reflects the efficiency of the combined human-AI team performance. (0% - 100%). Measured directly from user input (was the solution modified? Yes/no), translated into % across the operator's multiple interactions with AI-generated solutions
Significance of human revisions	The extent of human revisions compared to the AI decision. Here, small, localized revisions (e.g., merging two small adjacent sectors in the northeast corner of the FIR) would be rated differently from larger or multiple revisions across various areas in the FIR.	Reflects the AI system performance. (LOW, MED, HIGH interaction %). Measured directly from user input (of the modified solutions, how much interaction was measured? LOW number and extent of changes, MEDIUM number, and extent of

		changes HIGH number and extent of changes), translated into % across the operator's multiple interactions with AI-generated solutions
System Reliability	System trustworthiness - operation as expected under several conditions without major failures.	Reflects the efficiency of the combined human-AI team performance. (0%-100%). Measured directly from how many times the AI-generated solutions are sound or lead to failures
AI prediction robustness	How accurately and robustly does the AI system predict a certain sectorisation over a certain time horizon. Does re-evaluation of the sector structure in a shorter time horizon lead to different solutions? It is undesirable if small variations in capacity lead to significant differences in the sector structures/routings.	Reflects the efficiency of the combined human-AI team performance. Measured directly from the AI generated solutions. How big a variation in capacity has to be to cause the AI to revise its previous solutions.
Prompt demand rate	Assess how many times the ATCO prompts additional explanations from the AI generated solutions.	Reflects the AI system performance. (LOW, MED, HIGH interaction %) Measured directly from user input (how much interaction with explanations occurred and how the generated scenario is rated using the 'dynamic density index', measuring complexity), translated into % across the operator's multiple interactions with AI-generated solutions
AI co-learning capability	Does the ATCO feel that by the end of the trial runs, the AI has learned his preferences?	Links to the desired output of the AI system. (Likert, 7-points scale).
Human Response Time	Needed response time to react to AI advisory/information.	(LOW, MED, HIGH response time %). Measured directly from user input (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), translated into % across the operator's multiple interactions with AI-generated solutions.
Reduction in Delays	Percentual reduction of flight delays due to AI implementation in airspace and air traffic management.	0% - 100%
Workload perception	Assess ATCOs perception of the system impact on their workload (either positive or negative).	Likert, 7-points scale1 (Huge Increase in workload) to 7 (Huge decrease of workload)

**1.6 Features of use case**

<b>Task(s)</b>	Planning, prediction, optimization, interactivity, recommendation.
<b>Method(s)</b>	Supervised Learning (e.g., ensemble decision trees) and Reinforcement learning.
<b>Platform</b>	<a href="#">BlueSky</a> digital environment.



### 1.7 Standardization opportunities and requirements

<i>Classification Information</i>
<b>Relation to existing standards</b>
<p><b>ISO/IEC 23894:2023</b>, Information technology — Artificial intelligence — Guidance on risk management. Autonomous management and optimization of sectorisation in pre-tactical ATM operations are high-stake tasks, and therefore, risk management specifically related to AI is fundamental.</p> <p><b>ISO/IEC 38507:2022</b>, Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations. Autonomous AI requires an analysis of governance implications and also a redefinition of the organization structure.</p> <p><b>ISO/IEC 24029-2:2023</b>, Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for using formal methods. Since artificial neural networks can be a component of the autonomous AI system, formal methods to assess the robustness properties of neural networks are fundamental to certify and monitor autonomous systems.</p> <p><b>ICAO DOC 4444</b> – Standards and Recommended Practices in Air Traffic Management</p> <p><b>ERNIP Part 3</b> – EUROCONTROL Procedures for Airspace Management, Airspace Management Handbook for the Application of the Concept of the Flexible Use of Airspace.</p> <p><a href="https://www.sesarju.eu/masterplan2020">https://www.sesarju.eu/masterplan2020</a> - European ATM Master Plan</p>
<b>Standardization requirements</b>
<p>Establish a standard set of KPIs for measuring the performance of AI-based airspace structuring systems, and how the AI performance compares to heuristic methods in prediction and planning systems.</p>

### 1.8 Societal concerns

<i>Societal concerns</i>
<b>Description</b>
<p><b>Increased air traffic density in Europe:</b> The challenge of maintaining safe and efficient air traffic management under increased traffic loads while adhering to the workload capacity limits of tactical ATCOs.</p> <p><b>Privacy and data protection:</b> The use of AI in ATM airspace structuring (routing and sectorisation) involves the collection and analysis of large volumes of data, including potentially sensitive information. There is a concern about how data is stored, processed, and protected, especially in compliance with data protection regulations like GDPR.</p> <p><b>Transparency and accountability:</b> There is a societal demand for transparency in how AI systems make decisions, especially in high-stake transportation systems like ATM. The public might be concerned about the lack of understanding of AI decision-making processes and the accountability mechanisms in place in case of failures or errors.</p> <p><b>Employment and skill shift:</b> The full automation of the airspace structuring (routing and sectorisation) tasks might lead to concerns about job displacement and the need for reskilling of ATC staff. While AI can optimize operations, it also changes the nature of work, requiring a shift in skills for human operators who now need to oversee and interact with advanced AI systems.</p> <p><b>Public trust and acceptance:</b> For the successful implementation of AI in air transportation, gaining and maintaining public trust is crucial. There may be apprehensions and resistance from the public regarding the shift to AI-driven systems, especially among those accustomed to traditional methods.</p>
<b>Sustainable Development Goals (SGD) to be achieved</b>
<p>SGD9. Industry, innovation and infrastructure / SGD11. Sustainable cities and communities / SGD13. Climate action</p>

## 2 Environment characteristics

<i>Data characteristics</i>	
<i>Observation space</i>	<p>Partially observable.</p> <p>Data update is near real-time with a certain look-ahead time (minutes up to hours).</p> <p>Domain: defined on a continuous space.</p> <p>Size: &gt; 2000 flights per day, with &gt; 10 observable states per flight, &gt; 8 sectors with &gt; 20 coordination points (entry and exit points) per sector</p> <p>Noise: The observation can be noisy due to unsynchronized update frequencies and data quality of various data platforms (e.g., weather updates).</p>
<i>Action space</i>	<p>Mixed action space: sectorisation decisions are discrete (e.g., ‘split’ and ‘merge’), but sector geometry can vary on a continuous space depending on the algorithmic approach. Routing decisions are continuously characterized by waypoint locations. The action space of a human ATCO (for routing advisories) is three-dimensional (altitude, heading, speed).</p> <p>Size: The action space of the human ATC staff manager is limited to the number of sectors to choose from and depends on ATCo staff availability, the number of flights, and the weather conditions (determining geographic restrictions). The action space of the human ATCO is three-dimensional (altitude, heading, and speed) and depends on the number of flights in the sector.</p> <p>Time horizon: sectorisation and routing actions range typically from a few minutes to a couple of hours (= pre-tactical operations)</p>
<i>Type of task</i>	<p>Human staff managers and AI assistants act in a sequential environment: the previous decisions can affect all future decisions. The next action of these agents depends on what action they have taken previously and what action they are supposed to take in the future.</p>
<i>Sources of uncertainty</i>	<p>Stochastic (weather forecasts, variability in traffic load, unpredicted ATCo staff shortage, variability in opening and closing MIL areas)</p>
<i>Environment model availability</i>	<p>Yes (aircraft performance models, ISA standard atmosphere)</p>
<i>Human-AI interaction</i>	<p>Co-learning between the human and AI: AI assistant proposes a sectorization and routing plan, the human staff manager and planner ATCO evaluates the plan, and human agents accept or revise the plan (= feedback to AI assistant).</p>

### 3 Technical details

#### 3.1 Actors

<i>Actor Name</i>	<i>Actor Description</i>
FMP supervisor	<p>The human FMP supervisor is responsible for implementing a sectorisation plan and routing structure on a pre-tactical time scale. The FMP supervisor needs to evaluate the outputs of an AI assistant that aims to support the supervisor in generating sectorisation and routing suggestions.</p>
AI assistant	<p>The AI assistant provides sectorisation plan and routing suggestions to the FMP supervisor. It takes predicted information about the environment from various systems (e.g., weather forecasts from METEO services, traffic loads from Central Flow Management Unit, ATCo staff schedule, etc.) and historical data. In the training phase, it can act on the environment to evaluate its recommendations. In the evaluation/testing phase, the actions on the environment should be performed by the human only.</p>

Environment	The FMP supervisor interacts with the BlueSky digital environment and with the AI assistant through a secondary interface. The AI assistant can also portray its sectorisation and routing recommendations directly in the BlueSky environment (top-down Earth map).
-------------	--

## 4 Step-by-step analysis of use case

### 4.1 Overview of scenarios

Scenario conditions					
No.	Scenario name	Scenario description	Triggering event	Pre-condition	Post-condition
1	Nominal operational conditions	The condition is used as a baseline, allowing the comparison of minimum KEA routings devised by the AI system under nominal operational conditions with routings devised in restricted airspace availability conditions. Traffic loads over a typical day (24 hours) will be used as inputs.	Nominal traffic load over 24 hours, including periods of inbound and outbound of Lisbon FIR.	Nominal ATCO staffing capacity. Normal weather conditions.	The system proposes and/or executes efficient flight routes and sectorisation plans and presents results on an auxiliary interface for the human supervisor to evaluate. These results are then used as a baseline for comparison with scenarios with restricted airspace availability.
2	Military restrictions	This scenario deals with decreased airspace availability due to the activation of one or two military areas. Traffic should be routed around the military-restricted airspace while minimizing the KEA and adhering to sector capacity limits, which may require off-standard sectorisation.	Activation of one or two military areas.	Nominal traffic load over 24 hours. Nominal ATCO staffing capacity. Normal weather conditions.	The system proposes and/or executes efficient flight routes and off-standard sectorisation and presents results on an auxiliary interface for the human supervisor to evaluate.
3	Environmental disturbances	This scenario deals with highly decreased airspace availability due to challenging weather conditions, reducing the availability of airspace on a short time horizon.	Challenging weather conditions.	Nominal traffic load over 24 hours. Nominal ATCO staffing capacity. No active military areas.	The system proposes and/or executes efficient flight routes and off-standard sectorisation and presents results on an auxiliary interface for the human supervisor to evaluate.
4	Large perturbation	This scenario deals with decreased airspace availability	Activation of more than two military areas in	Nominal ATCO staffing capacity.	The system proposes and/or executes efficient flight routes and off-standard

		due to the activation of more than two military areas, in conjunction with challenging weather conditions, further reducing on a short time horizon the availability of the airspace. This case simulates an edge-case situation.	conjunction with challenging weather conditions.		sectorisation and presents results on an auxiliary interface for the human supervisor to evaluate.
--	--	---	--	--	--

#### 4.2 Steps for all scenarios

For each scenario the number of steps are the same and in-line with current practices in capacity flow & management and sectorisation on medium- to long-term time scales.

Step no.	Event	Name of process/ activity	Description of process/ activity Service	Information producer (actor)	Information receiver (actor)	Information Exchanged
1	Start	The FMP supervisor prepares his/her shift	FMP supervisor selects a maximum time horizon for a sector plan and enters that information into the system. The shift is prepared taking into account the forecasted traffic, the airspace restrictions, and the available ATCOs	FMP supervisor	AI assistant	SET
2	Initialise plan	AI assistant generates an initial plan	The FMP supervisor requests an initial sectorisation and routing structure from the AI assistant. This includes portraying a horizontal and vertical sector layout on a map and/or secondary interface, a network of KEA efficient routings, a timeline showing ATCo staff and traffic occupancy per sector, and a time slider enabling the FMP supervisor to preview changes in sectorisation and routings 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.	AI assistant	FMP supervisor	SRPLAN

3	Plan evaluation	The FMP supervisor evaluates the plan	The AI assistant may propose several alternative sector plans and routing structures, each with a different probability values (based on historical data), KEA efficiency scores, and robustness scores depending on ATCO and traffic capacity, fluctuations in predicted traffic load, and uncertainty in weather forecasts. Using the time slider, the FMP supervisor can evaluate the probability, efficiency, and robustness scores for different times within the maximum look-ahead time horizon.	AI assistant	FMP supervisor	STATE
4	Human interacts	The FMP supervisor interacts with the plan	The FMP supervisor interacts with the suggested sector plan and routings 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 sector merge or split and adjust one or two traffic streams); 3) revise large sections of the plan (e.g., revise multiple sectorisation events across various time horizons and revise several traffic streams).	Staff manager	AI assistant	DEC
5	Re-schedule	Trigger an alert to re-schedule	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 and routing structure, the AI assistant issues an alert, triggering the FMP supervisor to go back to Step 2.	AI assistant	FMP supervisor	AL

## 5 Information exchanged

<b>Information exchanged (ID)</b>	<b>Name of information</b>	<b>Description of information exchanged</b>
SET	Inputs and settings for AI assistant	FMP supervisor sets maximum time horizon for the AI assistant
SRPLAN	Sector plan	AI assistant suggestions for sectorization and routings.
STATE	Predicted system state	Predicted system state over a certain time period, including traffic load, weather conditions, ATCo capacity, sector and routing topology, probability, efficiency, and robustness scores.
DEC	Human decision / interaction with the AI assistant operator	FMP supervisor's choice in terms of accepting, nudging, and revising.
AL	AI assistant alert	AI assistant issuing an alert, signaling to the FMP supervisor that data used for predictions have changed significantly, warranting re-scheduling.

## 6 Requirements

<b>Requirements</b>		
<b>Categories ID</b>	<b>Category name for requirements</b>	<b>Category description</b>
Ro	Robustness	It encompasses both its technical robustness (the ability of a system to maintain its level of performance under a variety of circumstances) as well as its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no harm can occur unintentionally. <i>Source: EU-U.S. Terminology and Taxonomy for Artificial Intelligence. First Edition</i>
E	Efficiency	The ability of an AI system to achieve its goals or perform its tasks with optimal use of resources, including time, computational power, and data.
I	Interpretability	Make the behavior and predictions of AI systems understandable to humans, i.e., the degree to which a human can understand the cause of a decision. <i>Source: Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.</i>
Re	Regulatory and legal	The AI system's capacity to meet its objectives while complying with relevant laws, regulations, and ethical standards.
O	Other	Other non-function requirements related to environmental concerns and maintenance
<b>Requirement R-ID</b>	<b>Requirement name</b>	<b>Requirement description</b>
Ro-1	System resilience to unexpected events	The AI system should work correctly under a variety of conditions and withstand operational



		disruptions. This includes resilience to unexpected events like adverse weather and sudden changes in the ATCO staff availability.
Ro-2	Cyber and data security	Focuses on protecting the system against unauthorized access, cyber threats, and data breaches. This ensures the integrity and confidentiality of sensitive operational data and safeguards the system from malicious attacks.
Ro-3	The system's reliable operation and decisions	Shall show the capacity to perform its required functions under stated conditions for a specified period. This includes maintaining consistent performance and minimizing system failures or errors.
E-1	Capability to optimize resources and operations	The system shall maximize airspace and ATCO staffing utilization.
E-2	Scalability	Concerns the system's ability to handle growth in traffic loads, such as increased air traffic or airspace expansion, without performance degradation. This ensures the system remains effective as the scale of ATM operations increases.
I-1	Provide clear, understandable explanations for its decisions	It is crucial for human operators to validate and trust the AI's decisions, especially in restricted airspace conditions with complex sectorisation scenarios.
I-2	Usability of the system from the human and other stakeholders perspective	It should include intuitive interfaces, ease of use, and effective communication of information.
Re-1	Compliance with legal standards and regulations	Adherence to data protection laws, safety regulations, and ethical guidelines governing AI systems in public transportation and the EU AI Act.
O-1	Maintainability	Involves the ease with which the system can be maintained and updated. This includes the ability to diagnose and fix issues, update software, and adapt to changing operational requirements.
O-2	Environmental Sustainability	Addresses the system's impact on the environment. This includes considerations such as energy efficiency of the AI algorithms and the broader ecological footprint of the system's implementation and operation.

## 7 Common Terms and Definitions

Common Terms and Definitions	
Term	Definition
Air Traffic Controller (ATCO)	Human operator is responsible for directing air traffic through a volume of airspace in a safe (i.e., maintaining separation standards) and efficient manner (i.e., expediting the flow of traffic, reducing delays, and avoiding inefficiencies in flow track miles).
Air Navigation Service Provider (ANSP)	An organization that provides the service of managing the aircraft in flight or in the maneuvering area of an airport and which is the legitimate holder of that responsibility. In this use case, NAV Portugal is the considered ANSP.
Flight Information Region (FIR)	A three-dimensional area in which aircraft are usually under the control of a single authority (ANSP). Sometimes, one or more FIRs



	have a combined upper area control, and/or FIRs are vertically split into lower and upper sections.
Airspace sector	A three-dimensional geographical area within an FIR is under control by a single ATCO or multiple ATCOs (e.g., planner and executive controller). A FIR is commonly divided into multiple sectors.
General Air Traffic (GAT)	All aviation traffic conducted in adherence to the International Civil Aviation Organisation (ICAO) regulations.
Flow Management Position (FMP)	ANSP Unit responsible for sector capacity and traffic flow management

# ANNEX 3 – RELEVANT ALTAI REQUIREMENTS

## POWER GRID

### REQUIREMENT #1 Human Agency and Oversight

#### Human Agency and Autonomy

The end-users are fully aware that the decision comes from an AI-based system. The AI assistant is designed to provide recommendations to human operators in managing the power grid, which, in case of failure, might endanger the safety of property and people, lead to electricity outages, and affect humans and the economy. However, human operators remain in charge of implementing actions.

The assistant may create addictive behavior in humans, but in normal conditions (i.e., without an adversarial attack to the output), it will not manipulate user behavior. However, with time, humans may start to trust more in AI, and there is the risk of over-reliance. Technically, this issue is solved by the requirement of alarms from the AI assistant when it cannot provide a recommendation. These alarms should be designed carefully, as the AI can generate confusion via multiple actions and too many alarms.

#### Human Oversight

The AI system provides recommendations that the human can accept or adapt at will. The human can override the AI system when necessary. Humans already know the type of output (i.e., the same as traditional tools in power system control rooms). Still, operators should be trained to understand the rationale behind the AI system (e.g., understanding how RL works) and its limitations. The alarm is issued when the AI system cannot generate a recommendation that effectively overcomes the problem, e.g., lack of knowledge, or high uncertainty. Moreover, it can leverage simulation (with a physically-based tool – power flow) to understand the impact of each recommendation in the system.

### REQUIREMENT #2 Technical Robustness and Safety

#### Resilience to Attack and Security

The system can endanger property safety and people or outages in the electrical grid. This can occur due to different reasons:

- Cyberattacks to input data, AI model output, and AI model
- Noise and missing input data
- High epistemic (model) uncertainty due to a lack of training data.

The security requirements of the UCs cover the potential sources and forms of cyber-attacks. The metrics to monitor the robustness of the AI system during training and operation should be defined for development. Red team/pen test, measures to ensure the integrity, robustness, and overall security of the AI system against potential attacks over its lifecycle are important for the final product but are out-of-scope for this project.

### General Safety

The following threats were identified during the analysis of UCs:

- The state vector (i.e., characterization of the operating context) that might have missing data, gross errors, or even adversarial attacks, which may lead to wrong decisions and could be classified as an environmental threat. The risk of possible malicious attacks should be evaluated. Misuse of the AI system can lead to wrong decisions from the human operator.
- Exposure to weather events.

Changes in the learned (pre-trained) AI system performed by supervised and reinforcement learning algorithms should be auditable and controlled by humans through automatic mechanisms to detect data and model shifts.

### Accuracy

In this concept of AI assistant, humans remain in control. The main consequence of low-level accuracy could be distrust from humans in AI and algorithmic aversion. Human control prevents critical consequences that could occur if low-accuracy AI actions were implemented automatically.

Transfer learning and adaptability of AI are important properties for deployment in real-world operating conditions. The second UC covers cases when the system operates in a different environment than the one used for training the AI.

Continuous monitoring of the AI system is fundamental. It should be done at different levels:

- Measure performance continuously (online) with metrics such as reward score (objective function), human operator acceptance rate, alarms utility function, and KPIs defined in the UC. This performance can be quantified both during training and operational phases (e.g., identify changes in the environment compared to the training phase).
- Stress tests will be conducted to assess the robustness of the AI system, considering perturbation in the state vector (input data). These tests may also consider perturbations in the model (e.g., weights) and output.

### Reliability, Fall-back Plans and Reproducibility

The reliability of the AI system is defined not only by its technical performance but also by its credibility with the human operator. An AI assistant with low reliability can recommend decisions that may not solve contingency problems and/or increase the risk of cascading effects. However, it will mainly lead to low trust from human operators and not direct adversarial or damaging consequences. In case of predictions with low confidence, the AI system generates alarms that inform the human operator.

Reproducibility is important when it is necessary to justify certain decisions to the grid stakeholders (e.g., Energy regulator, curtailed renewable energy produces), which means that the same decisions should be obtained using the original data, AI model and code.

Governance procedures should be defined to re-train (or conduct maintenance over) the AI assistant in case of continuous poor performance.

Verification and validation methods are required and will be proposed in WP4.

### **REQUIREMENT #3 Privacy and Data Governance**

#### Privacy

Privacy concerns are not an issue in the UCs.

#### Data Governance

The project's data management plan covers the measures required according to the GDPR; however, no personal data is being used during training.

The AI system may leverage historical records of actions taken by human operators (i.e., imitation learning), which is fully anonymized since the operator's identification is not required. Yet, the action's timestamp is required, and when cross-referenced with a table of operator shifts, it may be possible to identify the operator and corresponding actions (and performance).

### **REQUIREMENT #4 Transparency**

#### Traceability

In power grids, traceability is fundamental, and transmission system operators keep historical records of all main events. Thus, it is possible to replay scenarios where AI was used. This means also storing the AI model (e.g., artificial neural networks weights, hyperparameters) together with the input and output data.

#### Explainability

Explainability is an important target of the projects and the developments. However, in this stage, most approaches rely on neural networks, where only feature importance (from sensitivity analysis or Shapley values) can be derived.

#### Communication

In this case, the human operator knows that an AI system is giving recommendations.

An alarm system for the AI system is foreseen – this is related to the concept of meta-awareness of AI-assistants that is discussed in the conceptual framework (see section 3.2.2.2.2). The goal is to inform the user when the AI system may fail to solve the technical problem. This alarm can be generated with information about the operating context (using the input data/state vector as raw information) and the model uncertainty (epistemic uncertainty). Corresponding situations shall be evaluated in the second UC (Sim2Real).

The AI4REALNET digital environments can be used within a training programme for operators on how to use and interact with the AI system.

### **REQUIREMENT #5 Diversity, Non-discrimination and Fairness**

#### Avoidance of Unfair Bias

It is prohibited for the system to favor specific producers of energy unfairly. A level playing field in the energy market, as well as fair competition, must be provisioned. Measures must be implemented to ensure these fairness constraints are observed.

In the discussion, the two following issues regarding this proposition emerged:

1. Occurring bias may originate from technical or physical limitations of electrical grid operations and hence may (in part or wholly) not be avoidable.
2. Requiring the AI system to adhere to fairness standards that are not required from existing alternative techniques may put it at a disadvantage, especially if those originate from the source of the previous issue.

In these UCs, bias and discrimination can be directed towards certain grid users (generators, flexible loads) that are redispatched (or curtailed) more frequently than others. Using the physical equations of the power grid, it is possible to compare the decisions made by the AI system and the impact that other grid users would have in solving the technical problem. For instance, ex-post, it is possible to run an optimal power flow (OPF) with the redispatch costs and compare its solution with the AI system. Having a least-cost solution is the primary goal.

#### Accessibility and Universal Design

The power grid operation is concerned with providing electricity to its customers. This objective is not influenced by the variety of preferences and abilities in society.

#### Stakeholder Participation

Stakeholders have been consulted during UC design and can be involved in the AI system design. Competitions with the digital environments will also help understand the AI system's benefits, limitations, and risks and extract lessons for further improvement.

### **REQUIREMENT #6 Societal and Environmental Well-being**

#### Environmental Well-being

The AI system will prioritize carbon-free actions, e.g., changing network topology to avoid renewable energy curtailment. A KPI for carbon intensity is considered in the UC. The AI system will also increase resilience to extreme weather events and reduce the cost of blackouts.

#### Impact on Work and Skills

The AI system will augment human operator analytical capabilities and decision-making tasks. It is not intended to replace the human.

Human operators in control rooms already use supporting tools (mainly classical tools) to develop and validate their decisions. However, a higher knowledge of the fundamentals behind the AI system can help human operators understand the decision-support process, and the proper use of data-driven tools requires training programs and risk assessment methodologies for humans and organizations.

#### Impact on Society at large or Democracy

Not relevant for these UCs.

## REQUIREMENT #7 Accountability

### Auditability

A third-party audit is unlikely during the development phase. During the operational phase, an audit might occur in case of outages, blackouts, or cyber-attacks on the input data.

Saving the AI model (weights, hyperparameters, structure) is essential for the auditability and traceability of the recommendation of the AI assistant.

AI Act will likely demand an audit (high-risk system). If the audit is to be repeated often, it may become necessary to develop an automated procedure.

### Risk Management

A process for third parties to report potential vulnerabilities, risks, or biases in the AI system is fundamental, particularly the creation of a database similar to [AVID](#). A recent initiative in this direction is the [AI Risk Repository](#) from MIT. However, the vulnerabilities and risks of other systems (e.g., SCADA) should be evaluated together due to interdependencies with the AI system (e.g., source of input data).

## RAILWAY

## REQUIREMENT #1 Human Agency and Oversight

### Human Agency and Autonomy

The AI system is designed to interact with and guide human end-users and make decisions that affect humans and society. It directly impacts human autonomy, has the potential to generate overreliance, and can negatively affect or manipulate the end-user's decision-making process. It is therefore, important that when an operative version of the technology is developed, the implementation partner ensures that the employees using the tool are trained to do so and are made aware that they are interacting with an AI system.

The planned system does not simulate social interaction and should, therefore, not risk creating human attachment. However, like in the power grid, it can stimulate addictive behavior.

### Human Oversight

The systems developed for UC2.Railway include a self-learning and partially autonomous agent, with human oversight ranging from Human-in-the-Loop to Human-in-Command. It is important for employees who use such a tool to be properly trained. In cases where the AI system is the executing agent, a procedure must be in place with which operations can be safely transferred back to full human control.

Step 8 of the process described in UC2.Railway in Annex 2 pertains to "Human review and system adjustment", which should include at the least the following:

- Detection and response mechanisms for undesirable adverse effects of the AI system for the end-user

- Oversight and control measures to reflect the self-learning or autonomous nature of the AI system

## **REQUIREMENT #2 Technical Robustness and Safety**

Technical robustness and safety can only be considered to a certain degree within the scope of the AI4REALNET project. These questions must be re-considered when the developed solutions are implemented in operations. It can generally be said that there is little to no risk of physical harm to humans or damage to infrastructure/material, as collision avoidance in railway systems is handled by a separate system, which limits the impact of threats from technical faults, defects, outages, attacks, misuse, inappropriate and malicious use.

### Resilience to Attack and Security

A requirement was defined in the UC to ensure the developed systems are compliant with relevant safety and security standards. However, certification, security coverage, and procedures ensuring the integrity, robustness, and overall security of the AI system against potential attacks over its lifecycle are out of scope for the AI4REALNET project.

### General Safety

Potential risk areas and metrics can be identified during development. However, concrete assessment of risk levels and evaluation of risk metrics is out-of-scope. Safety nets, fault tolerance, and technical robustness/safety review depend on the practical implementation of the technologies developed. This being said, Step 8 of the UC considers a regular review of the system by a human expert.

### Accuracy

There is potential for negative societal/financial impacts resulting from low system accuracy. Step 8 of the UC includes monitoring and documentation of the system's accuracy.

When implementing the developed solution in real-world operations, it is important for the system's performance to be continuously monitored and documented and for employees to be informed on expected accuracy levels.

### Reliability, Fall-back Plans and Reproducibility

A detailed analysis of the system's risks and damaging consequences in case of low reliability is out of the scope of this project, as the system will be developed in a simulated environment. However, the validation and verification methods of reliability, as well as failsafe fallback plans, will be considered. It is specifically required that the transition from algorithmic to full human control is pollable at any time and that the operator is explicitly notified when the system yields uncertain results or predictions with low accuracy.

Online learning can lead to unforeseen changes in behavior, necessitating explainability and interpretability requirements for the continuous learning process. To ensure that continuous learning does not interfere with the system's reliability, it is required to document the online learning process and make it interpretable for humans, allowing for continuous monitoring.

### **REQUIREMENT #3 Privacy and Data Governance**

#### Privacy

Requirements ensure compliance with legal standards and regulations. The further mechanisms to distinguish and flag privacy concerns should be considered at later stages of development, but they are out of the scope of this project.

#### Data Governance

The system does not use private data for training or in a productive environment. The measures according to the GDPR are out of scope for this project, as they are not relevant to the types of data used.

### **REQUIREMENT #4 Transparency**

#### Traceability

Traceability is fundamental to keep historical records of all main events. Thus, it is possible to replay scenarios where AI was used. This means also storing the AI model together with the input and output data.

#### Explainability

Explainability and interpretability are essential for human operators and supervisors. Interpretability requirements ensure that agent goals, option generation, decision-making, and learning are transparent and understandable to the human agent.

#### Communication

The system is designed as a software tool, ensuring that it is always clear to human agents that they are interacting with an artificial agent. Human operators and supervisors must understand the capabilities and limitations of the AI system to prevent misuse and foster trust. The requirements ensure that all aspects of the AI system (agent goals, option generation, decision-making, learning, capabilities, and limitations) are communicated to the human agent.

### **REQUIREMENT #5 Diversity, Non-discrimination and Fairness**

#### Avoidance of Unfair Bias

Bias avoidance must be considered during development and monitored after implementation. Fairness requirements are introduced to ensure that the system fairly distributes unavoidable delays throughout the system and does not unfairly favor specific Railway Undertaking Operating Managers (RUOMs). Analysis of end-user groups and diversity considerations cannot be done for a proof-of-concept. Also, out of scope for this project but potentially interesting for further development are mechanisms to detect and flag issues related to bias and discrimination.

#### Stakeholder Participation



Incorporating stakeholders in the design process ensures that the developed systems fit real-world requirements. Workshops involving both stakeholders and the public were conducted at the early stages of development planning and will continue parallel to the development of the algorithm.

## **REQUIREMENT #6 Societal and Environmental Well-being**

### Environmental Well-being

While the AI system is not expected to directly impact the environment, an improved system efficiency may result in a positive environmental impact.

### Impact on Work and Skills

As the system will impact human work and work arrangements, the potential impacts of the developed systems must be understood ahead of time so that design considerations can be made during development. Workshops are recommended to receive feedback from the intended end users as well as psychological considerations provided by Human Factors experts, both of which inform and guide the development process of the PoC. The new work arrangements will require some new skills, so the design and realization of training courses are essential for implementation, albeit out-of-scope for a PoC.

## **REQUIREMENT #7 Accountability**

### Auditability

Requirements ensure retrospective quality control. Documentation and logging ensure auditability and is essential for post-hoc analyses and performance evaluations.

### Risk Management

Overall, risk analysis and training are out-of-scope for a POC, but planned documentation and logging build a foundation with which to establish monitoring mechanisms for internal assessment of AI ethics and accountability of the system.

## **REQUIREMENT #1 Human Agency and Oversight**

### Human Agency and Autonomy

AI systems can generate confusion if the prediction is not within the expected outcome. Still, such a situation cannot have an operational impact, as the operator will have the final decision. The end-users are aware that they are interacting with an AI system; additionally, the decision should be communicated through a separate platform for an additional visual reminder of the decision's origin.

No risk of addiction or manipulation is expected according to the current description of the UCs. However, indirectly, the use of an assistant and the higher acceptance rate of decisions can, for a

longer time, affect the confidence and awareness of operators and reduce the time and effort they invest in checking the decisions generated by the assistant.

It is planned to change the system from only generating recommendations for human revision and approval or adjustment to fully automated at the later development stages. The transition from a lower to a higher level of automation would affect human autonomy and demand stricter rules.

### Human Oversight

The level of oversight will change with the higher level of autonomy, and the “management by exception” system becomes more autonomous. At a lower level of autonomy, the operator checks each decision and only implements it if it is considered safe. The manual check by the operator does not apply when the decisions are implemented automatically. To ensure no undesirable effects, concrete requirements can be derived from KPIs: system reliability and AI prediction robustness.

An alarm is issued under two conditions:

- The AI system generates an alarm to the human operator when it cannot produce a recommendation that solves the problem, and the human operator must decide. This is related to the concept of meta-awareness discussed in the conceptual framework (see section 3.2.2.2.2).
- An environmental change can affect the generated decision’s validity. The operator must review the decision under new circumstances.

## AIR TRAFFIC MANAGEMENT

### REQUIREMENT #2 Technical Robustness and Safety

#### Resilience to Attack and Security

At a higher level of automation, if the decision of the system leads to dangerous situations and is implemented without the need for human confirmation, this can lead to damage.

Security requirements demand that the system be protected against unauthorized access, cyber threats, and data breaches.

#### General Safety

Stability and reliability are essential in an AI assistant. Two robustness requirements specify that the system should work correctly under normal and unexpected circumstances. To ensure this, risk evaluation is essential for the design of the safety properties of the systems. After evaluation, the risks should be included in training materials.

As algorithms are based on online RL, changes should be logged in the model or algorithm design, and clear notifications should be given to the operator if the version of the system changes during a decision process. This will help avoid confusion if the new version exhibits different behavior.

#### Accuracy

Although the system only serves as a recommender, if it produces decisions with a low level of accuracy, there is still a risk of adverse consequences in the case of a deficient performance of the human-in-the-loop operator.

It is mentioned in the UC description that the data is coming from many sources, and the data infrastructure is demanding. A validation procedure is advisable to ensure the correct database is used.

The KPIs acceptance and agreement score are based on comparing AI-generated suggestions and decisions accepted by the operator. These are good metrics to ensure the accuracy of the AI system.

#### Reliability, Fall-back Plans and Reproducibility

At a low level of autonomy, human operators will monitor the decisions and estimate the risks before applying an AI-generated recommendation. At a higher level of autonomy, automatically implemented decisions can lead to adverse consequences. To evaluate and ensure different aspects of the AI system's reliability and reproducibility, such KPIs as the significance of human revisions, system reliability, and AI prediction robustness can be logged for continuous monitoring and analysis of the system performance.

Governance procedures should be developed to specify the conditions for fallback; this is specifically important for the later development stages, when the decisions are implemented automatically, without the need for consent from the human operator.

### **REQUIREMENT #3 Privacy and Data Governance**

#### Privacy

No private data is planned to be used during training or operation.

#### Data Governance

The calculation of these KPIs will involve the processing of personal data, which must be fully anonymized. Since the identification of individuals who made or revised decisions is irrelevant for these calculations, all data must be handled in a way that ensures anonymity, protecting personal information while preserving the accuracy of the KPIs.

### **REQUIREMENT #4 Transparency**

#### Traceability

Currently, information is scattered over various ATM systems, which makes the oversight of the input data and their quality assessment even more important for the accuracy of the decisions. It is planned to log all human interventions into AI decisions. The logging can be extended by the documentation of input data that were used to generate the decision.

#### Explainability

Humans can consult additional information and explanations underpinning AI decisions on demand, which is expected to foster trust in the system’s decisions and acceptance of the AI system.

The KPI “Trust in AI solutions” score describes the operator's confidence in the AI-generated solution, with and without the need for additional explanations. Evaluating the difference between the KPI with and without explanation can show if the explanations are helpful. The KPI “Prompt demand rate” shows how often the operator needs additional explanation and what kind of explanations are used, which can serve as a source of information about potential system improvements. A routine for continuous surveys can be implemented as a part of the assessment of human-system interaction, the result of which can be logged together with the implemented decision.

#### Communication

During the training process, operators will be informed about the purpose, criteria, and limitations of the decision(s) generated by the AI system.

### **REQUIREMENT #5 Diversity, Non-discrimination and Fairness**

#### Avoidance of Unfair Bias

The biases are not expected in the UC description.

#### Accessibility and Universal Design

The AI will provide recommendations to human operators. It will not directly impact target end-users and/or subjects.

The ATC staff may be impacted by the AI system regarding their workload. While AI can optimize operations, it also changes the nature of work, requiring a shift in skills for human operators who now need to oversee and interact with advanced AI systems.

The introduction of the AI system might lead to concerns about job displacement and the need to reskill ATC staff.

#### Stakeholder Participation

Stakeholders have been consulted during UC design and can be consulted during the AI system design. Surveying operators will also help understand the AI system's benefits, limitations, and risks and extract lessons for further improvement.

### **REQUIREMENT #6 Societal and Environmental Well-being**

#### Environmental Well-being

The system aims to reduce the load on the air traffic system and reduce the environmental impact by reducing carbon emissions. Metrics can be developed to calculate the saved carbon emissions and the system’s positive environmental impact.

#### Impact on Work and Skills

The AI system will augment human operator analytics capabilities and decision-making tasks. Training prior to the implementation of the new AI system should help the operator overcome their doubts or fears concerning the change in their work methods. Extended knowledge about the fundamentals behind the AI system can help human operators understand the decision-support process. Training should be provided before the AI system is implemented.

*Impact on Society at Large or Democracy*

No impact is expected.

**REQUIREMENT #7 Accountability**

*Auditability*

Traceability of the recommendation of the AI assistant down to the model should be ensured. Saving the AI model (weights, hyperparameters, structure) and input data is essential for auditability.

Due to the nature of RL algorithms, continuous learning will change the state of the system and should be audited after each system update.

*Risk Management*

The code for the models developed during the project will be made publicly available as baselines for future benchmarking efforts, with reproducibility being a central requirement. This approach ensures that other researchers and practitioners can trace and verify the results. It is important to clarify that AI4REALNET does not anticipate the operational deployment of these AI systems in a real-world environment. Should such deployment become a consideration, it would be essential to develop robust software methodologies to ensure the traceability and accountability of the operational algorithms.

# ANNEX 4 – CONTEXT, CHARACTERISTICS, IMPACT AND EVALUATION OF DECISIONS

This annex details the elements of section 3.1.2.

## WORD ANALYSIS

### METHODOLOGY

Context, characteristics, impacts, and evaluation of the decision process contain a list of questions and answers per domain, allowing for a characterization from each domain’s perspective.

For each question, the list of the most significant words has been extracted: it corresponds to the set of words that are put in bold. For example, the list of words for characteristics of decision for the “Time constraints” category and Air Traffic domain is the following: “Strategic planning,” “Operational adjustments,” “Unexpected conditions,” “Pre-tactical,” and “Tactical.”

Then, a similarity analysis is performed between each pair of items of different domains within a given category by calculating a cosine similarity<sup>36</sup> between embedding<sup>37</sup> of the corresponding items. A threshold (e.g. 0.8<sup>38</sup>) above which 2 items from 2 different domains are considered “similar” is defined.

The score of similarity is the percentage share of “similar” pairs of items when crossing 2 domains, compared to the total number of pairs of items (this can be computed at different levels, e.g., per category, per domain, etc.).

### RESULTS

This sub-section presents the result of the similarity analysis performed between each pair of items of different domains for each category: context, characteristics, impacts, and evaluation of decisions.

#### SIMILARITY ANALYSIS FOR CONTEXT OF THE DECISION PROCESS

Category	Air Traffic-Electricity	Electricity-Railway	Railway-Air Traffic
<b>Constraints</b>	exceeding network		
	capacity, network	-	-
	capacity		

<sup>36</sup> Normalized dot product of X and Y, see, for example, the cosine\_similarity from the scikit-learn library (<https://scikit-learn.org/stable/index.html>)

<sup>37</sup> Embedding is calculated using the sentence-transformers library and the open-source models from Huggingface platform. Models used for embedding are picked from the sbert.net sentence transformer library (<https://www.sbert.net/index.html>) and, more specifically, from pre-trained semantic search models ([https://www.sbert.net/docs/sentence\\_transformer/pretrained\\_models.html#semantic-search-models](https://www.sbert.net/docs/sentence_transformer/pretrained_models.html#semantic-search-models)) trained on scientific citations and can be used to estimate the similarity of two publications (SPECTER).

<sup>38</sup> Corresponds to the 80 percentile.

Category	Air Traffic-Electricity	Electricity-Railway	Railway-Air Traffic
<b>Forecasting</b>	-	operational delays, traffic flow and congestion, outage risk, topology, infrastructure failure risks	weather impact, weather condition
<b>Observations</b>	information from different platforms, external context, availability of actions, external events	power grid state, external context,  signal and control system status, weather conditions, availability of actions, external events	external events
<b>Operators</b>	multiple operators, one or multiple operators	multiple operators	multiple operators, one or multiple operators
<b>Possible Events</b>	-	-	health emergencies, regional or global health emergencies, technical failures, environmental conditions, adverse weather, operational disruptions, accidents and emergencies, operational disruptions
<b>Uncertainty</b>	maintenance operations, human factors, technical failures, outage planning	maintenance operations, technical failures, human factors, outage planning	technical failures, human factors, weather and environmental conditions

#### SIMILARITY ANALYSIS FOR DECISION CHARACTERISTICS

Category	Air Traffic-Electricity	Electricity-Railway	Railway-Air Traffic
<b>Action Type</b>	preventive or corrective	preventive or corrective, preventive (operational adjustments)	preventive or corrective, preventive (operational adjustments)
<b>Implementation</b>	planned or real-time	real-time, planned or real-time	real-time, planned or real-time
<b>Size of action space</b>	large and mixed action space	large and mixed action space	large and mixed action space

Category	Air Traffic-Electricity	Electricity-Railway	Railway-Air Traffic
<b>Time Constraints</b>	-	-	operational adjustments, strategic planning, unexpected conditions, tactical, maintenance scheduling
<b>Time step</b>	real-time to medium-term, real-time to long-term	real-time to long-term	real-time to medium-term, real-time to long-term
<b>Trade-offs</b>	-	risk versus consequences, cost versus innovation, safety versus efficiency	operational flexibility versus standardization, cost versus innovation, capacity versus quality of service, safety versus efficiency, capacity versus efficiency

#### SIMILARITY ANALYSIS FOR IMPACTS OF A DECISION

Category	Air Traffic-Electricity	Electricity-Railway	Railway-Air Traffic
<b>Recovery time</b>	-	actions can be reverted in a couple of hours, can vary significantly, depending on several factors	-
<b>Lasting Effects</b>	-		-

#### SIMILARITY ANALYSIS FOR EVALUATION OF A DECISION

The evaluation is based on KPIs that have been defined for each UC and grouped into the following categories:

- Technical quality of AI-based solutions,
- Quality of AI-based solutions as perceived by human operators,
- Human-AI interaction,
- Efficiency of combined human-AI performance,
- Cognitive load,
- Robustness,
- Trustworthiness.



Category	Air Traffic-Electricity	Electricity-Railway	Railway-Air Traffic
<b>Cognitive load</b>	workload perception, workload, assistant disturbance, human response time	workload, assistant disturbance, human information processing	workload perception, human response time, human information processing
<b>The efficiency of combined human-AI performance</b>	-	total decision time, response time, ability to anticipate	-
<b>Human-AI interaction</b>	-	decision support for the human operator, human learning, human control and autonomy over the process	-
<b>Quality of AI-based solutions perceived by human operators</b>	efficiency score, assistant relevance, agreement score	assistant relevance, acceptance, situation awareness, comprehensibility	decision support satisfaction, comprehensibility, significance of human revisions, prompt demand rate, acceptance, assistant relevance, efficiency score, acceptance score, agreement score
<b>Robustness</b>	AI prediction robustness, generalization to different grids	-	-
<b>Technical quality of AI-based solutions</b>	network utilization, reduction in delays	network utilization, topological action complexity, delay reduction efficiency, punctuality	reduction in delays, response time, delay reduction efficiency
<b>Trustworthiness</b>	trust towards the AI system, trust in AI solutions score	human motivation, trust towards the AI system	trust towards the AI system, trust in AI solutions score

## **DETAILED ANSWERS**

The chapters hereafter list the detailed answers about context, characteristics, impacts, and evaluation of decisions for each domain.

All detailed answers have been used to extract the main characteristics of decisions across domains.

## **CONTEXT OF THE DECISION PROCESS**

Hereafter, a list of questions allows for characterizing the context in which the decision process is conducted regarding following subtopics: network and resource management, event handling and uncertainty.

## NETWORK AND RESOURCE MANAGEMENT

### Network Capacity: What defines network capacity?

Air Traffic	Electricity	Railway
<p><i>NM = Network Manager (EUROCONTROL Unit)</i> <i>CNS = Communication; Navigation; Surveillance</i> <i>ATCO = Air traffic controller</i></p> <p>The network capacity depends on:</p> <ul style="list-style-type: none"> <li>• <b>Area (airspace dimension)</b></li> <li>• <b>Route structure</b></li> <li>• <b>Aeronautical (CNS) Systems/Equipment availability</b></li> <li>• <b>Demand (NM)</b></li> <li>• <b>Airport Infrastructure</b></li> <li>• <b>Staff availability (ATCO):</b> For instance, when planning long/mid-term resources, it is essential to ensure that the training requirements do not overly detract from the availability of air traffic controllers for real-time operations.</li> </ul>	<p>The overall network capacity depends on the <b>capacity of each of its given transmission lines</b>, the latter being defined by:</p> <ul style="list-style-type: none"> <li>• A maximum current threshold</li> <li>• A maximum duration during which the amount of current flowing on the line can reach this threshold.</li> </ul> <p><i>In the context of the project, the capacity of the network is only considered from the point of view of its capacity to pass current. Other issues, such as voltage, inertia, or stability, are not considered in the scope of AI4REALNET.</i></p>	<ul style="list-style-type: none"> <li>• <b>Train Frequency and Schedule</b></li> <li>• <b>Operational Strategies:</b> The approach to managing train operations, including prioritization of certain types of trains</li> </ul> <p><i>In the context of the project, following elements have less priority.</i></p> <ul style="list-style-type: none"> <li>• Track Layout and Infrastructure</li> <li>• Signal and Control Systems</li> <li>• Train Length and Composition</li> <li>• Maintenance and Upkeep: Regular maintenance ensures that all components of the railway infrastructure are in optimal condition, reducing the likelihood of failures that can decrease capacity</li> </ul>

**Constraints: What are the constraints or congestion issues?**

Air Traffic	Electricity	Railway
<p>Capacity constraints arise from unpredictable events that affect the nominal sector capacity:</p> <p><b>Network capacity</b> (see above)</p> <p><b>Military area activation</b></p> <p><b>Adverse weather or disruptive events</b></p> <p><b>Separations</b> (between aircrafts)</p> <p><b>Sectorization</b> (to increase capacity)</p> <p><b>ATCO Workload</b></p>	<p><b>Congestion arises when network capacity is exceeded.</b> It can occur in:</p> <ul style="list-style-type: none"> <li>• “N” situation, where all elements of the transmission grid are available (not considering planned maintenance)</li> <li>• “N-k” situation, in case of an unplanned outage (k = 1 in practice, except for very specific cases)</li> </ul> <p>Congestions can arise because of too much generation in some area that needs to get evacuated (especially with new renewable plants and high wind or PV gradients), too much consumption in some area that needs to be supplied without local generation, high imports or exports to other countries, or weaker transmission grid because of planned or unplanned outages. Storms, Snow, and Big National events can be days with higher congestions in real-life.</p> <p><i>Note: other types of constraints can arise, such as voltage constraints, dynamic stability, etc., that are not considered in the scope of the project.</i></p>	<p><b>Train Frequency and Schedule Constraints:</b> High demand for track access by various types of trains (freight, passenger, high-speed) can lead to scheduling conflicts and congestion. The rigid scheduling of trains can reduce the system's ability to adapt to real-time demands or disruptions.</p> <p><b>Operational Strategies:</b> Fixed prioritization of certain types of trains (e.g., passenger over freight) can lead to suboptimal utilization of network capacity. Inflexible operational strategies may not adequately respond to varying demand or unexpected disruptions.</p> <p><b>Human Factors:</b> Resistance to change, skill gaps, and concerns about job security among railway staff can pose significant challenges to the adoption of AI technologies and the transition to more autonomous systems.</p> <p><i>In the context of the project, following elements have less priority.</i></p> <p><b>Track Layout and Infrastructure Constraints:</b> The physical limitations of the railway network, such as track layout, tunnels, bridges, and</p>

Air Traffic	Electricity	Railway
		<p>crossings, can limit the capacity and flexibility of train operations. Tight curves, steep gradients, and bottleneck points can restrict train speed and frequency.</p> <p><b>Signal and Control System Limitations:</b> The efficiency and reliability of signal and control systems directly impact train movement and safety. Outdated or malfunctioning signal systems can cause delays, reduce track capacity, and limit the potential for automation and real-time decision-making.</p> <p><b>Train Length and Composition Issues:</b> Longer trains can carry more cargo or passengers per trip but may face restrictions on certain routes due to platform lengths and track layouts. The composition of trains (e.g., mixed freight) also affects handling and speed, potentially leading to inefficiencies and congestion.</p> <p><b>Maintenance and Upkeep:</b> Inadequate or irregular maintenance can lead to equipment failures, track damage, and signal issues, directly impacting capacity and safety. The challenge is to balance maintenance needs with operational demands, minimizing</p>

Air Traffic	Electricity	Railway
		<p>downtime while ensuring infrastructure integrity.</p> <p><b>Integration with Other Modes of Transport:</b> The inability to efficiently integrate railway operations with other modes of transport (e.g., road, maritime) can create bottlenecks at transfer points, affecting the overall efficiency of cargo and passenger flows.</p> <p><b>Regulatory and Safety Constraints:</b> Regulatory requirements, including safety standards and operational guidelines, can impose limitations on train operations, affecting scheduling, train composition, and the adoption of new technologies or operational strategies.</p> <p><b>Technological and Data Limitations:</b> The effectiveness of AI-based solutions is heavily dependent on the availability and quality of data. Limitations in data collection, processing, and sharing can hinder the development and implementation of AI algorithms.</p>

**Operators: Are one or multiple operators involved?**

Air Traffic	Electricity	Railway
<p><b>Several stakeholders</b> are involved in a Collaborative Decision-Making process:</p> <ul style="list-style-type: none"> <li>• Systems Technical Supervision (STS) &amp; Maintenance staff CNS/ATM</li> <li>• ATCO (air traffic management)</li> <li>• Airlines</li> <li>• Airport operators</li> <li>• EUROCONTROL Network Manager (NM)</li> <li>• National Air Force (FAP)</li> </ul> <p>In terms of the air traffic management task, generally, <b>the decision is taken by one sole operator</b> (supervisor or tactical ATCO).</p> <p>If not a real-time decision, <b>there may be more contributors for the decision under the lead of one operator</b> (e.g., planner ATCO, FMP supervisor).</p>	<p>Depending on the complexity of the problem, <b>multiple operators</b> might need to coordinate to make the decision: other RTE’s control center, field operators, market participant operators, DSO operators, etc.</p> <p>One operator is always leading: this is defined ex-ante according to operational rules (e.g., in case a line crosses several RTE’s control center areas, one control center is leading, or in case of escalation needs, operations’ management people can be involved).</p> <p><i>In this project, we will consider only one given operator managing congestion and interacting with the AI.</i></p>	<p>The railway system requires the collaboration of <b>multiple operators</b>, encompassing those managing infrastructure, train operations, maintenance, and integration with other transport modes.</p> <p>This multi-operator environment is necessary to address the diverse constraints and ensure efficient, safe railway operations, particularly when integrating AI technologies.</p>

## EVENT HANDLING AND UNCERTAINTY

### Possible Events: What kind of events can happen?

Air Traffic	Electricity	Railway
<p><b>Partial airspace closure</b></p> <p><b>Operational disruptions:</b> system failures, staff strikes (leading to deviations), corrective maintenance</p> <p><b>Adverse Weather</b></p> <p><b>Sector Capacity overload</b></p> <p><b>Cybersecurity incidents</b></p> <p><b>Accidents and emergencies</b></p> <p><b>Regional or global health emergencies</b> like Pandemics or regional health crises can affect staff availability and demand for travel, forcing the need to adapt operations.</p>	<p><i>In this project, only events and uncertainties applicable to operations conducted in control centers and related to network and constraints as defined above are considered.</i></p> <p><b>Outage of a transmission grid element</b> (e.g., transmission line), which can have several causes:</p> <ul style="list-style-type: none"> <li>• Planned maintenance for several days</li> <li>• Unplanned maintenance (a list of non-urgent operations is updated, and in case operating conditions are favorable, they are carried out)</li> <li>• Unplanned outages due to unplanned events (controlling device failures or any other failure, storm, thunderstorm, malicious acts, human factor)</li> </ul> <p>The list of unplanned outages is predefined and continuously monitored: in case of severe anticipated weather conditions (e.g., thunderstorm), additional outages are monitored (such conditions can lead to outages that would be less probable in normal conditions).</p>	<p><b>Technical Failures:</b> This includes signal failures, malfunctioning control systems, and breakdowns of trains or infrastructure components. Technical failures can lead to significant delays and safety risks.</p> <p><b>Environmental Conditions:</b> Extreme weather conditions such as heavy snowfall, storms, floods, or landslides can damage infrastructure, obstruct tracks, and lead to service disruptions.</p> <p><b>Accidents and Emergencies:</b> Collisions, derailments, or incidents at level crossings can have severe consequences for safety, service continuity, and infrastructure integrity.</p> <p><b>Operational Disruptions:</b> These can arise from unexpected maintenance issues, power outages, or failures in communication systems, impacting the regular flow of train services.</p> <p><b>Health Emergencies:</b> Pandemics or localized health crises can affect staff availability,</p>



Air Traffic	Electricity	Railway
	<p>In the case of important and strategic <b>infrastructure projects</b>, outages can be planned years or even multiple years in advance.</p> <p>The outage can also concern <b>generation units</b> (even if these assets are not under TSOs’ direct responsibility): maintenance planning of generation units is important (especially if the unit is large).<i>In this project, following elements are modeled: Planned maintenance with a standard duration of about a day and without delays, Unplanned events such as line contingencies and unexpected line disconnections occur. A standard duration of several hours is applied for recovering the asset.</i></p> <p><b>Differences between forecasted flows and real flows:</b> TSOs are simulating flows on their transmission grid to anticipate as best as possible the consequences of possible events. This requires relevant grid modeling and forecasts, as well as a definition of monitored events (according to operational policies). Thus, important discrepancies between forecasted flows and real flows are possible events that can worsen the operational conditions. Such events can originate from:</p>	<p>demand for travel, and necessitate changes in operations to ensure passenger safety.</p> <p><i>In the context of the project, following elements have less priority.</i></p> <p><b>Cybersecurity Incidents:</b> Attacks on the control systems, data breaches, or disruptions to operational technology can pose significant risks to safety and operations.</p> <p><b>Human Factors:</b> Errors or misjudgements by staff, strikes, or other labour-related issues can affect train scheduling and operational efficiency.</p> <p><b>Regulatory Changes:</b> New safety, operational, or environmental regulations may require adjustments in operations, sometimes on short notice.</p> <p><b>Demand Fluctuations:</b> Sudden increases or decreases in passenger or freight demand, possibly due to external events or seasonal variations, can strain capacity and scheduling.</p> <p><b>Integration Challenges with Other Modes:</b> Delays or disruptions in other transport modes can have a knock-on effect on railway</p>

Air Traffic	Electricity	Railway
	<ul style="list-style-type: none"> <li>• Bad inputs from market participants</li> <li>• Actions or failures from market participants (e.g., a generation unit is not generating as planned)</li> <li>• Forecasting errors for highly stochastic weather-based sources such as renewable energy sources or demand</li> <li>• Bad modelling in the tools, incorrect hypothesis, delayed outages</li> </ul> <p><i>In this project, stochasticity from local generation and demand is considered. Market players bids are not given explicitly but forecast for generation are provided.</i></p>	<p>operations, especially at intermodal transfer points.</p>

**Observations: What are the important observations for perception?**

Air Traffic	Electricity	Railway
<p><b>Information</b> from several other platforms:</p> <ul style="list-style-type: none"> <li>• NM (sector capacity and demand)</li> <li>• Adherence to the plan (departure from airports)</li> <li>• Area availability</li> <li>• CNS equipment’s availability (e.g. planned maintenance)</li> </ul> <p><b>External events</b> that will impact ATM (regarding airport congestion and airplanes parking): Summits (e.g., Web summit), major events (Soccer tournament’s finals, music festivals)</p>	<p><b>Power Grid state</b>, which is composed of:</p> <p>Information needed to estimate the usage level of network capacity and remaining margin:</p> <ul style="list-style-type: none"> <li>• Measures</li> <li>• Topology (e.g., circuit breaker positions)</li> </ul> <p>Information needed to elaborate the forecasted grid modelling:</p> <ul style="list-style-type: none"> <li>• Topology</li> <li>• Generation schedules and forecasts (some of which are linked with weather forecasts)</li> <li>• Demand forecast</li> <li>• Maintenance planning (with associated criticality)</li> <li>• Localization of demand and RES generation</li> <li>• Market conditions (in some cases, this can help improving RES forecasts)</li> </ul> <p><b>Availability of actions:</b> operators must know the current state of flexibilities and the one that are currently available. For instance, some lines or substations might need some cooldown time before being switched again, or generation</p>	<p><b>Infrastructure Condition:</b> Monitoring the physical condition of tracks, bridges, tunnels, and other critical infrastructure components for signs of wear, damage, or other issues that could impact safety or performance.</p> <p><b>Weather Conditions:</b> Observing weather patterns and environmental conditions that could affect railway operations, such as temperature extremes, precipitation, wind speeds, and natural disasters.</p> <p><b>Train Status and Performance:</b> Tracking the real-time status of trains, including their location, speed, and operational health. This also involves monitoring the condition of onboard systems and components.</p> <p><b>Signal and Control System Status:</b> Keeping tabs on the functionality and performance of signal systems and automated control mechanisms to ensure they are operating correctly and safely.</p> <p><b>Traffic Flow and Congestion:</b> Observing the movement of trains throughout the network to identify bottlenecks, congestion, or</p>

Air Traffic	Electricity	Railway
	<p>units can have activation time constraint to consider for redispatching.</p> <p><b>External context information</b> that can imply</p> <ul style="list-style-type: none"> <li>• Additional outage risks (due to storm, thunder, fires)</li> <li>• Putting extra attention (e.g., important events such as Olympics),</li> <li>• Modifying the operation (e.g., important accidents, ongoing fires, or protests) on some parts of the grid.</li> </ul> <p>This means that operators are liaising with other authorities.</p> <p><i>In this project, external context information is not considered.</i></p>	<p>potential conflicts that could lead to delays or safety risks.</p> <p><b>External Events:</b> Being aware of events or situations outside the railway system that could impact operations, such as road traffic conditions near crossings, public events affecting demand, or disruptions in other transport modes.</p> <p><i>In the context of the project, following elements have less priority.</i></p> <p><b>Passenger and Freight Volumes:</b> Monitoring passenger flows and freight loads to manage capacity effectively and anticipate demand fluctuations.</p> <p><b>Maintenance and Upkeep Activities:</b> Tracking scheduled and unscheduled maintenance activities to ensure they are completed efficiently and do not unduly disrupt operations.</p> <p><b>Regulatory and Compliance Changes:</b> Keeping updated on changes in regulations or safety standards that could affect operational practices or require adjustments to equipment or procedures.</p>

Air Traffic	Electricity	Railway
		<p><b>Cybersecurity Threats:</b> Monitoring for signs of cyber threats or vulnerabilities that could impact operational technology, control systems, or data integrity.</p> <p><b>Human Factors:</b> Observing the performance and behavior of personnel involved in railway operations to identify potential errors, inefficiencies, or areas for improvement in training and operations.</p>

**Uncertainty: What can be delayed or uncertain?**

Air Traffic	Electricity	Railway
<p>Accidents/incidents impact in airports, such as:</p> <p><b>Weather and Environmental Conditions:</b> Predicting weather conditions and their impact on operations involves a degree of uncertainty. Extreme or adverse weather can cause unexpected delays or the need to apply changes in operations.</p> <p><b>Technical Failures:</b> The occurrence of technical failures, such as systems breakdowns (although there exists redundancy in aeronautical systems, some failures can impact infrastructure availability are inherently unpredictable, and may impact operations and cause delays.</p> <p><b>Human Factors:</b> Variability in human behavior, such as operator errors or unexpected absences, can introduce uncertainties into the system.</p> <p><b>Potential Cybersecurity Threats:</b> although the probability of a cyber-attack is very low due to the defensive barriers, if it happens, it will have a huge impact on operations.</p> <p><i>In the context of the project, following elements have less priority.</i></p>	<p>Uncertainty concerns:</p> <ul style="list-style-type: none"> <li>• <b>Transit flows</b> on grid elements (transmission lines) because of generation or consumption uncertainties, as well as unexpected event</li> <li>• <b>Possibility to carry out a given remedial action</b> (the associated uncertainty is normally very low thanks to operational rules, procedures, maintenance, and operation preparation)</li> </ul> <p>Delay can impact:</p> <ul style="list-style-type: none"> <li>• <b>Maintenance operations</b> on the grid</li> <li>• <b>Outage planning</b> from generation units (especially large ones)</li> </ul>	<p><b>Infrastructure Repairs and Upgrades:</b> The timing and duration of infrastructure maintenance or upgrade projects can be uncertain, affected by unforeseen complications or delays in securing necessary materials or approvals.</p> <p><b>Weather and Environmental Conditions:</b> Predicting weather conditions and their impact on railway operations involves a degree of uncertainty. Extreme weather can cause unexpected delays or necessitate changes in operations.</p> <p><b>Technical Failures:</b> The occurrence of technical failures, such as signal system malfunctions or rolling stock breakdowns, is inherently unpredictable, leading to unplanned delays and operational disruptions.</p> <p><b>Human Factors:</b> Variability in human behavior, such as operator errors or unexpected absences, can introduce uncertainties into the system. Additionally,</p>

Air Traffic	Electricity	Railway
<p>Technical Infrastructure &amp; Upgrades: The timing and duration of infrastructure maintenance or even upgrade projects can be uncertain and affected by unforeseen complications or delays.</p>		<p>the response time and decision-making process in emergency situations can vary.</p> <p><b>Passenger and Freight Demand:</b> Fluctuations in passenger numbers and freight volumes can be uncertain and affected by factors like economic conditions, public events, or changes in consumer behavior.</p> <p><i>In the context of the project, the following elements have less priority.</i></p> <p><b>Regulatory Changes:</b> The timing and impact of regulatory changes, including new safety or environmental regulations, can be uncertain, requiring adjustments to operations or equipment that may not have clear implementation timelines.</p> <p><b>Traffic Congestion:</b> Predicting traffic flows and congestion levels on the railway network involves uncertainty, particularly in the face of disruptions, special events, or sudden changes in demand.</p> <p><b>Supply Chain Delays:</b> Uncertainties in the supply chain, affecting the availability of spare parts for maintenance or upgrades, can lead to delays in scheduled works or repairs.</p>

Air Traffic	Electricity	Railway
		<p><b>Cybersecurity Threats:</b> The nature and timing of cybersecurity threats can be highly uncertain, with potential impacts on operational technology and control systems that are difficult to predict in advance.</p> <p><b>Integration with Other Transport Modes:</b> Delays or uncertainties in other transport modes, such as road or maritime transport, can impact railway operations, particularly at intermodal transfer points.</p>



Forecasting: What can be forecasted?

Air Traffic	Electricity	Railway
<p><b>Military airspace activations</b></p> <p><b>Weather condition</b></p> <p><b>Known ATCO staff shortages</b></p> <p><b>Scheduled maintenance</b> of CNS equipment</p>	<p><b>Transit flows</b> on grid elements (transmission lines), including all required inputs (forecasts, outages, etc.)</p> <p><b>Local generation or consumption</b> is still difficult; this is often more done at a regional or national level. <i>In this project, we will consider local forecasts for simplifications.</i></p> <p><b>Outage risk</b> due to weather conditions. However, an unplanned outage (of a line or generation unit) cannot be forecasted as such.</p> <p><b>Topology</b> can be forecasted in case actions are conducted as part of a long-term strategy.</p>	<p><b>Passenger and Freight Demand:</b> AI can predict fluctuations in passenger numbers and freight volumes based on historical data, seasonal trends, economic indicators, and special events, helping to optimize capacity and scheduling.</p> <p><b>Traffic Flow and Congestion:</b> AI can forecast potential congestion points and traffic flow by analyzing current and historical traffic patterns, enabling better resource allocation and operational planning.</p> <p><b>Weather Impact on Operations:</b> Advanced weather forecasting models can predict the impact of weather conditions on railway operations, allowing for preventive adjustments to schedules and maintenance plans.</p> <p><b>Infrastructure Failure Risks:</b> Predictive maintenance tools use data from sensors and historical maintenance records to forecast the likelihood of infrastructure or equipment failures, minimizing downtime and preventing disruptions.</p> <p><b>Operational Delays:</b> ML algorithms can analyze patterns in operational data to predict delays,</p>

Air Traffic	Electricity	Railway
		<p>identify root causes, and suggest improvements to reduce future occurrences.</p> <p><i>In the context of the project, the following elements have less priority.</i></p> <p><b>Service Disruptions:</b> AI can predict potential disruptions to service, including those caused by technical failures, environmental conditions, or external events, facilitating quicker response and mitigation strategies.</p> <p><b>Energy Consumption and Costs:</b> AI can forecast energy consumption and costs for train operations, assisting in optimizing energy use and identifying cost-saving opportunities.</p> <p><b>Safety and Security Risks:</b> By analyzing incident reports, security data, and external threat intelligence, AI can predict potential safety and security risks, enhancing preventative measures.</p> <p><b>Regulatory and Compliance Changes:</b> While challenging, AI systems can track regulatory trends and predict future compliance requirements, helping railway operators avoid legal changes.</p>

Air Traffic	Electricity	Railway
		<p><b>Technological Advancements and Adoption Rates:</b> Predictive analytics can estimate new technologies' impact and adoption rates within the railway sector, guiding investment and development decisions.</p> <p><b>Economic and Societal Trends:</b> AI can analyze broader economic and societal trends that might affect railway operations, such as shifts in urbanization, trade patterns, or travel preferences.</p>

## DECISION CHARACTERISTICS

From a general point of view, 3 types of decisions can be distinguished:

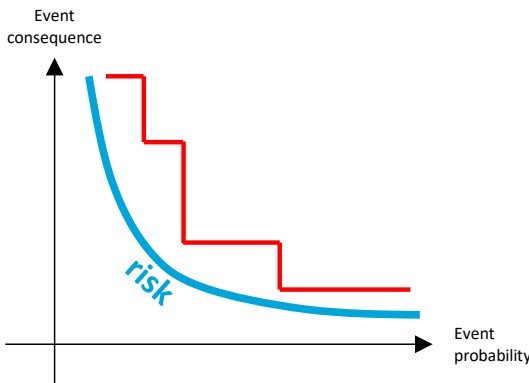
- Strategic Decisions: Long-term timeframe, large scope, high level of complexity, significant resources, and high impact.
- Tactical Decisions: Medium-term timeframe, moderate complexity, translate strategic goals into actionable plans, requiring analysis and coordination.
- Operational Decisions: Short-term timeframe, specific tasks or activities, low level of complexity, structured tasks with defined procedures.

Hereafter, a list of questions allows for characterizing more in detail the decisions taken, with following topics: tradeoffs, time constraints, time step, implementation, action type, size of action space.

**TRADEOFFS: WHAT ARE THE POTENTIAL TRADEOFFS (OR GOALS) INVOLVED?**

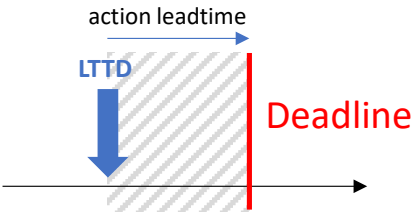
Tradeoffs allow for better characterization of the multiple objective nature of the decision process:

Air Traffic	Electricity	Railway
<p><b>Safety versus Capacity:</b> Higher capacity demand challenges keeping the desired safety level (the probability of occurrence of critical events increase).</p> <p><b>Capacity versus Efficiency:</b> As capacity increases, the efficiency rises if the routes are the desirable for each aircraft.</p> <p>When the existing conditions don't support the direct route as the most efficient, the aircraft pilot's preference/request cannot be satisfied. The efficiency interests have different points of view between aircraft crew (pilot) and ATC.</p>	<p>From a general point of view, real-time operation decisions shall search for the best compromise between:</p> <ul style="list-style-type: none"> <li>• personal safety</li> <li>• maintaining the operational safety of the System</li> <li>• compliance with the operating limits of the components of the electrical system</li> <li>• the quality of electricity and compliance with contractual commitments</li> <li>• optimization of interconnection capacities</li> <li>• cost reduction (congestion and losses)</li> <li>• guaranteeing the completion of maintenance work</li> </ul> <p>More specifically, risk management is centered around the <b>risk/consequences tradeoff</b>. The risk is defined by the product of:</p> <ul style="list-style-type: none"> <li>• <b>Consequences of the event</b>, assessed from a grid operation perspective (forecasts or close to real-time), which are, in the</li> </ul>	<p><b>Cost versus Innovation:</b> Investing in new technologies and innovations can significantly improve operations and customer satisfaction in the long term but may require substantial upfront costs and financial risk.</p> <p><b>Capacity versus Quality of Service:</b> Expanding capacity to accommodate more trains or passengers might strain resources or degrade service quality, affecting punctuality, comfort, and overall customer experience.</p> <p><b>Operational Flexibility versus Standardization:</b> Standardizing operations can lead to efficiencies and easier management but may reduce the system's ability to adapt to local conditions or unexpected disruptions.</p> <p><i>In the context of the project, following elements have less priority.</i></p> <p><b>Short-term Gains versus Long-term Sustainability:</b> Decisions may offer immediate improvements or cost savings but could undermine long-term sustainability goals, such</p>

Air Traffic	Electricity	Railway
	<p>context of this project, the constraints coming from the transit flows</p> <ul style="list-style-type: none"> <li>• <b>Probability of the event</b>, assessed mainly from <i>ex-ante</i> statistical studies performed on realized events</li> </ul> <p>This means that a low-probability event with important consequences will be considered similarly to a high-probability event with few consequences:</p>  <p>The above figure means that:</p> <ul style="list-style-type: none"> <li>• Events are classified on the x-axis according to predefined frequencies (e.g., once every 10 years)</li> <li>• A threshold is defined (red stepped curve) on the y-axis for each category of events,</li> </ul>	<p>as environmental impact reduction or infrastructure resilience.</p> <p><b>Customer Satisfaction versus Operational Constraints:</b> Enhancing customer satisfaction through more services, lower fares, or enhanced amenities may conflict with operational constraints, financial viability, or regulatory requirements.</p> <p><b>Maintenance versus Operational Availability:</b> Regular and thorough maintenance is crucial for safety and reliability but can reduce the availability of assets for operation, affecting service levels and financial performance.</p> <p><b>Innovation Adoption versus Workforce Impact:</b> Implementing automation and AI technologies can improve efficiency and safety but may lead to workforce displacement, requiring significant investments in retraining and change management.</p> <p><b>Data Privacy versus Operational Intelligence:</b> Collecting and analyzing data can significantly enhance operational intelligence and customer service but must be balanced against concerns about data privacy and security.</p>

Air Traffic	Electricity	Railway
	<p>with acceptable consequences (e.g., for rare events, it is possible to go up to a certain volume of load shedding)</p> <p>The blue curve depicts the possible risk assessment that can be performed in the operation context to refine the analysis.</p> <p><b>Consequences of the event</b> are ranked from acceptable to last resort, e.g.:</p> <ul style="list-style-type: none"> <li>• Maintenance planning delays and/or costs (especially for critical projects)</li> <li>• Involved costs</li> <li>• Load (or generation) shedding</li> <li>• Consequences on grid users</li> <li>• Size/volume of involved generation or consumption involved</li> </ul> <p>They also include side effects on other operational processes: redispatching must not hamper the balance of the system</p> <p><b>Probability of the event:</b> can be re-evaluated during operation: for example, an issue anticipated in a distant horizon might be less certain than an issue anticipated in 1h or less.</p>	<p><b>Regulatory Compliance versus Operational Flexibility:</b> Adhering to regulatory requirements ensures safety and standardization but can limit operational flexibility and the ability to adopt innovative solutions quickly.</p> <p><b>Environmental Impact versus Operational Needs:</b> Reducing the environmental impact of railway operations, for instance, through electrification or using alternative fuels, may require trade-offs with operational needs or cost implications.</p>

**TIME CONSTRAINTS: HOW MUCH TIME FOR DECISION-MAKING AND PLANNING?**

Air Traffic	Electricity	Railway
<p><b>Strategic planning:</b></p> <ul style="list-style-type: none"> <li>• Military areas activation (scheduled)</li> <li>• Planned reserved areas activation (balloons launch, etc.)</li> </ul> <p><b>Operational adjustments:</b></p> <ul style="list-style-type: none"> <li>• Military areas activation (unscheduled)</li> <li>• ATCO staff shortage</li> <li>• Sector capacity management</li> <li>• Airline/Pilot route adjustment request</li> </ul> <p><b>Unexpected conditions:</b></p> <ul style="list-style-type: none"> <li>• Adverse weather conditions</li> <li>• Unexpected airspace events (e.g., Volcanic ashes)</li> <li>• Inbound deviation of traffic flows from adjacent sectors/FIRs</li> </ul> <p><b>Pre-tactical</b> (up to 1-2 hours): sectorization</p> <p><b>Tactical</b> (few minutes): sectorization and traffic management</p>	<p>Each <b>decision to be taken in anticipation</b> is associated with its “LTTD” (Last Time To Decide), i.e., last moment to launch the action to that the effects are implemented before the targeted deadline.</p>  <p>If the action is launched after the LTTD, its effects will be implemented after the deadline</p> <p>LTTD can be calculated by subtracting the action lead-time from the targeted deadline.</p> <p><i>Note: indicative timings are defined on RTE’s side to allow considering duration associated to topological actions (e.g., up to 20 minutes to conduct 7 topological actions).</i></p>	<p><b>Emergency Responses:</b> In the case of emergencies or unexpected disruptions (e.g., accidents, technical failures), decision-making time is extremely limited. Decisions often need to be made in real-time or within minutes to ensure safety and minimize operational impact.</p> <p><b>Operational Adjustments:</b> Short-term operational decisions, such as rerouting trains due to a temporary obstruction or adjusting schedules in response to unexpected demand fluctuations, may have a slightly longer timeframe, ranging from minutes to a few hours.</p> <p><b>Maintenance Scheduling:</b> Decisions regarding routine maintenance and repairs often have a medium-term planning horizon. These decisions could be made days to weeks in advance, allowing for adequate preparation and resource allocation.</p> <p><b>Strategic Planning:</b> Long-term strategic decisions, such as infrastructure upgrades, the procurement of new rolling stock, or the implementation of significant technological innovations (like AI systems), involve extensive planning and consultation. The time frame for these decisions</p>



Air Traffic	Electricity	Railway
	<p>This concept also applies to decisions to be taken once an event has occurred. For example, when overload alarms are raised once the flows has exceeded a given threshold, LTTD are defined to act before the dilating effect on the cables becomes too dangerous and line is automatically switched off, meaning that these thresholds are defined such that associated LTTD is not zero</p> <p><b><u>Usual timeframes for decision making</u></b></p> <p>Considering a given delivery time T:</p> <p><b>Real-time:</b> T or after T</p> <p><b>Short-term:</b> from T-3h to T</p> <p><b>Mid-term:</b> one week to 3h before T</p> <p><b>Long-term:</b> several months before T up to one year</p> <p><b><u>Example of decisions depending on the different timeframes</u></b></p> <p><b>Real-time decisions:</b> implement a curative remedial action following an alarm for overload</p>	<p>can range from several months to years, given the need for detailed analysis, stakeholder engagement, regulatory approval, and financial planning.</p> <p><i>In the context of the project, the following elements have less priority.</i></p> <p><b>Regulatory Compliance and Safety Enhancements:</b> Decisions related to regulatory compliance or major safety enhancements may have varying time constraints, depending on the urgency of compliance deadlines or the critical nature of the safety issue. Planning and implementation could span from a few months to several years.</p> <p><b>Investment in Technology and Research:</b> Decisions to invest in research and development or to adopt new technologies for improving operations or customer service can also have a long lead time. The exploration, testing, and evaluation phases alone can take months or years before a decision on full-scale implementation is made.</p> <p><b>Capacity Expansion:</b> Decisions involving capacity expansion, such as adding new tracks and stations or expanding service areas, require extensive planning and are typically made within a horizon of several years. These decisions must account for</p>

Air Traffic	Electricity	Railway
	<p><b>Short-term or mid-term decisions:</b> implement a preventive remedial action for a forecasted constraint</p> <p><b>Long-term decisions:</b> make a contract with a producer to ensure the availability of a unit for dispatch (months or years in advance).</p>	<p>future demand projections, environmental impact assessments, and community engagement.</p>

**TIME STEP: WHAT IS THE TIME RESOLUTION FOR ANALYSIS?**

Air Traffic	Electricity	Railway
<p>Depends on the action:</p> <p>For <b>tactical phases</b> (flow management by tactical ATCO), it is real-time or a few minutes.</p> <p>For <b>strategic phases</b>, a short-medium-term analysis can be applied.</p>	<p><b><u>Real-time analysis</u></b></p> <p>The analysis is performed based on:</p> <ul style="list-style-type: none"> <li>• Grid models that have a high temporal resolution (up to 5min resolution) and represent the measured state of the system according to this resolution,</li> <li>• And/or real-time information from SCADA system, e.g., transit flows (usually 5s or 10s resolution).</li> </ul> <p><b><u>Short-term to Mid-term analysis</u></b></p> <p>Analysis is performed based on grid models that have an hourly resolution. It must be noted that such models are built using various data sources that have different time resolution, e.g.:</p> <ul style="list-style-type: none"> <li>• Generation schedules (5min resolution)</li> <li>• Interconnection exchange schedules (60min or 30min resolution)</li> <li>• Planned outages</li> </ul> <p><b><u>Mid-term to long-term analysis</u></b></p> <p>Grid models are created to represent a typical situation at a daily/monthly/yearly resolution.</p>	<p><b><u>Real-time or Near-real-time Analysis</u></b></p> <p><b>Operational Monitoring and Control:</b> For tasks like monitoring train locations, signaling status, or track conditions, data may be analyzed in real-time or near-real-time, often with time steps of seconds or minutes. This allows for immediate responses to operational changes or emergencies.</p> <p><b>Traffic Management:</b> Managing train movements and avoiding conflicts in densely trafficked areas require near-real-time analysis to make rapid adjustments.</p> <p><b><u>Short-term Analysis</u></b></p> <p><b>Service Performance Analysis:</b> Analyzing punctuality, service reliability, and passenger flow might be conducted daily or hourly to optimize scheduling and resource allocation for the following days.</p> <p><b>Maintenance Predictions:</b> Short-term maintenance needs, such as identifying equipment likely to fail or requiring servicing</p>

Air Traffic	Electricity	Railway
		<p>soon, may use analysis based on daily or weekly data.</p> <p><b><u>Medium-term Analysis</u></b></p> <p><b>Demand Forecasting:</b> Predicting passenger or freight demand to adjust services or plan for special events. Weekly or monthly data might be used to identify trends and adjust for upcoming periods.</p> <p><b>Resource Planning:</b> Planning for staffing, rolling stock availability, and maintenance schedules might be performed monthly, allowing for adjustments based on projected operational needs.</p> <p><b><u>Long-term Analysis</u></b></p> <p><b>Strategic Planning and Investment:</b> Decisions related to infrastructure investments, expansion plans, or long-term service changes may be based on analysis of trends and patterns identified in data spanning several months to years.</p> <p><b>Safety and Compliance Trends:</b> Analyzing safety incidents, compliance with regulations, and long-term performance trends might</p>

Air Traffic	Electricity	Railway
		utilize yearly data to inform policy adjustments and strategic safety initiatives.

**IMPLEMENTATION: IS DECISION REAL-TIME OR PLANNED?**

Air Traffic	Electricity	Railway
<p><b><u>Planned (pre-tactical or strategic phase)</u></b> In the case of programmed actions (e.g., <b>pre-known military areas activation</b>) or other constraints that are known in advance and enable planning measures.</p> <p><b><u>Real-time (tactical phase)</u></b> Redirecting traffic flows, either by airline request or for safety reasons.</p> <p>Operational adjustments derived from:</p> <ul style="list-style-type: none"> <li>• Sudden staff shortages (sickness, fatigue);</li> <li>• managing a sudden capacity overload due to any problem affecting an adjacent sector or FIR;</li> <li>• any emergency with immediate impact on traffic flows (e.g., aircraft in an emergency)</li> <li>• Reserved/restricted airspace activation</li> </ul>	<p>Congestion management often relies on a strategy, which is defined as a sequence of actions that will set the network topology for each timestep.</p> <p><b><u>Planned Implementation</u></b> In case the constraint is anticipated, and lead time for action is important, or a large risk can be mitigated.</p> <p><b><u>Real-Time Implementation</u></b> In all other cases, when flexibilities can be activated quickly</p>	<p><b><u>Real-Time Implementation</u></b> Real-time implementation is necessary when immediate action is required, usually in response to unforeseen events or to manage ongoing operations efficiently. This includes:</p> <p><b>Emergency Responses:</b> Implementing safety measures, rerouting trains, or adjusting operations in response to accidents, failures, or natural disasters.</p> <p><b>Operational Adjustments:</b> Making immediate changes to train schedules and routes or dispatching additional resources in response to real-time demand fluctuations or minor disruptions.</p> <p><b>System Monitoring and Control:</b> Continuous adjustments made by automated systems, such as signal control systems or AI-driven monitoring tools, to optimize performance and safety.</p> <p><i>In the context of the project, following elements have less priority.</i></p>

Air Traffic	Electricity	Railway
		<p><b><u>Planned Implementation</u></b></p> <p>Planned implementation is used for decisions that have long-term implications and require careful preparation, coordination, and resource allocation. This approach is characteristic of:</p> <p><b>Infrastructure Projects:</b> Expanding or upgrading tracks, stations, or signaling systems, which involves detailed planning, regulatory approvals, and significant investment.</p> <p><b>Strategic Initiatives:</b> Implementing new operational strategies, service expansions, or major technology overhauls, including the integration of AI systems for predictive maintenance or operational optimization.</p> <p><b>Maintenance Schedules:</b> Conducting routine or major maintenance work, which is planned to minimize impact on service and ensure resource availability.</p> <p><b>Policy Changes:</b> Implementing new regulations, safety protocols, or operational policies, which require training,</p>

Air Traffic	Electricity	Railway
		communication, and a phased approach to ensure compliance and effectiveness.



**ACTION TYPE: IS ACTION PREVENTIVE OR CORRECTIVE?**

The goal of this sub-section is to provide the most relevant examples in line with the use case description, keeping in mind what will be of importance within the AI4REALNET project.

Air Traffic	Electricity	Railway
<p><b><u>Preventive actions</u></b></p> <p>Planning the sectors merge/split in advance</p> <p><b><u>Corrective actions</u></b></p> <p>Re-directing (management) traffic flows</p>	<p>In principle, <b>all actions can be taken in a preventive or corrective manner</b>: the choice is made according to tradeoffs and relies on general characteristics of actions:</p> <ul style="list-style-type: none"> <li>• Availability of the action</li> <li>• LTTD,</li> <li>• Costs,</li> <li>• Etc. (see tradeoffs and impacts)</li> </ul> <p><i>In the context of the project, the actions considered can be as follows:</i></p> <p><b><u>Topological action</u></b></p> <p>These actions aim to redirect the energy flow on the power grid. It can be of two types:</p> <ul style="list-style-type: none"> <li>• switching on and off power lines between two substations</li> <li>• reconfiguring the busbar connection on a substation level. For instance, a “node splitting” changes the number of nodes from 1 node to 2 nodes in a substation</li> </ul> <p>All topological actions are discrete.</p>	<p><b><u>Preventive Actions</u></b></p> <p><b>Operational Adjustments:</b> Making changes to schedules, routes, or operational practices based on predictive models to avoid potential congestion, delays, or safety risks.</p> <p><i>In the context of the project, following elements have less priority.</i></p> <p><b>Routine Maintenance and Inspections:</b> Regularly scheduled checks and maintenance of tracks, rolling stock, and infrastructure to prevent failures.</p> <p><b>Predictive Maintenance:</b> Using AI and ML to analyze data from sensors and systems to predict equipment failures before they occur, allowing for targeted maintenance work.</p> <p><b>Training and Drills:</b> Conducting regular training sessions and emergency drills for</p>

Air Traffic	Electricity	Railway
	<p><b><u>Redispatching action</u></b></p> <p>This action aims at changing the power injection of a given flexibility (generator, load, battery, etc.) by adjusting the amount of generation output in the grid.</p> <p><b><u>Renewable energy curtailment</u></b></p> <p>Limits the power output of a given generation unit to a threshold.</p> <p><i>The following actions are listed for context but are not to be considered for the project:</i></p> <p><i>On RTE's side, in addition to topological actions decided and conducted by the operator, topological actions can be applied automatically by specific devices. The principle is that such devices monitor the flows on given lines and apply a predefined corrective action, possibly with priorities: for example, 1<sup>st</sup> topological change, then renewable energy curtailment.</i></p> <p><i>Actions can also include means to increase the number of available actions, for example:</i></p> <ul style="list-style-type: none"> <li><i>Delays or cancels planned maintenance (depending on consequences): This allows for more network switching/reconfiguration.</i></li> </ul>	<p>staff to ensure preparedness for potential incidents.</p> <p><b><u>Corrective Actions</u></b></p> <p><b>Repairs and Replacements:</b> Fixing or replacing faulty equipment or infrastructure components after a failure has been detected.</p> <p><b>Operational Recovery Plans:</b> Implementing contingency plans to recover from disruptions, such as rerouting trains, deploying replacement services, or adjusting schedules post-incident.</p> <p><b>Incident Investigations:</b> Conducting thorough investigations following accidents or failures to identify the root causes and implement measures to prevent similar incidents in the future.</p> <p><b>System Upgrades:</b> Upgrading or replacing systems and technologies found to be inadequate or prone to failure based on corrective feedback and incident analyses.</p>

Air Traffic	Electricity	Railway
	<ul style="list-style-type: none"><li>• <i>Establish contracts to ensure the availability of units for redispatching (usually done in the long-term for grid infrastructure projects)</i></li></ul>	

### SIZE OF ACTION SPACE: WHAT IS THE NUMBER OF POSSIBLE ACTIONS?

This Section is based on the description of the action space of use cases, and intends to show how complex the decision can be, especially from a “human only” perspective:

Air Traffic	Electricity	Railway
<p>The <b>size</b> of action space <b>depends on the three dimensions</b> defining the airspace size (lat-lon-altitude extents) and on the algorithmic approach.</p> <p>The action space of the human ATC staff manager is limited and depends on ATCO staff availability.</p> <p>The action space of the human ATCO depends on the number of flights in the sector.</p> <p>Action space is <b>mixed</b> (discrete and continuous).</p>	<p>The action space is <b>large</b>: e.g., for a network with around 100 nodes, it has more than</p> <ul style="list-style-type: none"> <li>• 65 000 different discrete actions,</li> <li>• 200 continuous actions</li> </ul> <p><i>For example, RTE’s grid is composed of more than 25 000 nodes and 10 000 lines.</i></p> <p>Action space is <b>mixed</b> (discrete and continuous).</p>	<p>While the action space grows linearly with the number of trains for the algorithmic part, it grows exponentially if there is a central actor controlling all the trains. The action space of the human dispatcher is, in any case, <b>exponentially growing with the number of trains</b>.</p> <p><i>For example, SBB operates more than 12,000 switches and 32,000 signals. In Germany, over 40,000 regional, long-distance, and freight train journeys take place every day.</i></p> <p>Furthermore, the dimensionality of the action space depends on infrastructure and timetable elements like switches, signals, and scheduled stops. Hereby, the impact on the dimensionality of the action space depends not only on the nature of the actor in the algorithmic part but also on the type of task, i.e., if the task is tackled episodically or sequentially on the algorithmic side.</p>

Air Traffic	Electricity	Railway
		Action space is <b>mixed</b> (discrete and continuous).

## IMPACT OF A DECISION

Hereafter, a list of questions allows for characterizing the impacts of the decision process and recovery.

### RECOVERY TIME: HOW MUCH TIME TO GET BACK TO NORMAL AFTER A DECISION?

Air Traffic	Electricity	Railway
<p>In ATM, the recovery time depends on the nature of the disruption.</p> <p>For instance, if the constraint is caused by adverse weather conditions or volcanic ashes, the recovery time depends on the closing of the abnormal occurrence.</p> <p>If the reason is due to military area activation, the time window is previously known and, therefore, controllable.</p> <p>In the case of an immediate change of conditions (rerouting a flight, ATCO staff shortage), there is no turning back to the previous condition.</p> <p>Offline time has a direct impact on recovery time.</p>	<p>In power grids, congestion must be relieved in a few minutes, otherwise, automatic protections are triggered (i.e., the line is automatically switched off) to avoid that, the problem amplifies and gets out of admissible bounds. Few actions might need to be coordinated in more complex cases.</p> <p>In general, <b>actions can be reverted in a couple of hours</b> at most after the overload conditions have vanished.</p> <p>However, this depends on the complexity of the event that triggered the actions or the consequences of the decision.</p> <p><i>Note: with regards to system balancing management, TSOs are supposed to balance the system within a 1h or 30min window, after all market trades have been performed.</i></p>	<p>The impact of a decision within railway operations and the subsequent recovery time can vary significantly based on the nature of the decision, the specific circumstances surrounding it, and the resilience of the railway system. Recovery time, or the duration required to return to normal operations following a disruption or implementation of a significant change, is influenced by several factors:</p> <p><b>Nature of the Disruption:</b> The type of event leading to a disruption (e.g., technical failure, environmental condition, accident) has a major impact on recovery time. For instance, recovering from a minor signal system glitch might take a few hours, whereas repairing damage from a severe weather event or an accident could take days or even weeks.</p> <p><b>Complexity of Operations:</b> The complexity of the railway network and its operations can affect recovery time. A highly interconnected system</p>

Air Traffic	Electricity	Railway
		<p>with dense traffic may take longer to recover, as disruptions can have cascading effects.</p> <p><b>Availability of Resources:</b> The availability of necessary resources, such as repair crews, replacement parts, and alternative transportation options, can significantly influence recovery time. Quick access to resources can expedite recovery, while shortages or delays can extend it.</p> <p><b>Effectiveness of Decision-making:</b> The effectiveness of the initial decision-making process, including the accuracy of forecasts and the efficiency of implemented mitigation strategies, plays a crucial role in determining recovery time. Decisions that effectively anticipate and address the core issues can lead to faster recovery.</p> <p><b>Resilience of Infrastructure:</b> The resilience of the railway infrastructure to withstand disruptions affects how quickly operations can normalize. Infrastructure designed with redundancy and quick repair capabilities can significantly reduce recovery time.</p> <p><b>Human Factors:</b> The response of personnel and their ability to adapt to and manage disruptions is critical. Effective training, clear communication,</p>

Air Traffic	Electricity	Railway
		<p>and strong leadership can enhance recovery efforts.</p> <p><i>In the context of the project, the following elements have less priority.</i></p> <p><b>Regulatory and Safety Requirements:</b> Compliance with regulatory and safety requirements can influence recovery time, as certain inspections and approvals may be necessary before normal operations can resume.</p> <p><b>Integration with Other Systems:</b> The ability to coordinate recovery efforts with other modes of transport and systems can affect the recovery time, especially for integrated transport networks.</p> <p><b>Preparedness and Pre-emptive Measures:</b> Systems that have pre-emptive measures in place, such as alternative routing plans and pre-staged resources, can recover more quickly than those that do not.</p> <p><b>Public and Stakeholder Communication:</b> Effective communication with passengers, freight customers, and stakeholders can mitigate the impact of disruptions and can be a crucial factor in the perceived recovery time.</p>



**LASTING EFFECTS: WHAT ARE THE POTENTIAL LASTING EFFECTS OF A DECISION?**

Air Traffic	Electricity	Railway
<p>In normal cases, lasting effects are the desired ones that led to the decision. The issues are solved, and the situation returns to normal.</p> <p><b>Bottleneck issues:</b></p> <ul style="list-style-type: none"> <li>delays in investments (upgrades, new systems/equipment)</li> <li>delays in regulation, lack of recruitment (to replace the retired staff, adaptation of operational staff)</li> </ul> <p><b>Punctuality</b></p>	<p>In normal cases, lasting effects are the desired ones that led to the decision, so, from a general point of view, the effect will be that congestion disappears and flows remain within <b>their acceptable limits</b>.</p> <p>Undesired consequences can be:</p> <ul style="list-style-type: none"> <li>Unanticipated side effects leading to <b>congestion on other parts of the neighboring grid</b></li> <li><b>Remedial action unavailable for other issues</b></li> <li><b>Damage to grid elements</b>, especially circuit breaker (the main intrinsic risk of using a circuit breaker is to damage it)</li> <li><b>Inability to perform planned outages</b> (or delays)</li> </ul> <p>The worst case is a decision leading to <b>transit flows exceeding admissible limits</b> as defined by operational policy (e.g., load shedding).</p> <p>Last, it can be possible in theory that decisions lead to threats to the security of people (including TSO’s staff): such a</p>	<p><b>Operational Efficiency:</b> Decisions that improve operational processes, such as the adoption of advanced scheduling systems or predictive maintenance, can lead to long-term improvements in efficiency, reducing delays and increasing the capacity of the railway network.</p> <p><b>Safety Enhancements:</b> Investments in safety technologies and practices can have a lasting impact on reducing accidents and incidents, enhancing the overall safety of the railway system for passengers and workers.</p> <p><b>Infrastructure Resilience:</b> Decisions to upgrade infrastructure or invest in more durable materials can increase the resilience of the railway network against environmental challenges, reducing the frequency and impact of disruptions.</p> <p><b>Customer Satisfaction:</b> Decisions that affect the quality of service, such as improvements in comfort, punctuality, and information provision, can have lasting effects on customer satisfaction and loyalty.</p>

Air Traffic	Electricity	Railway
	<p>consequence is the very first thing that any decision must avoid.</p>	<p><i>In the context of the project, the following elements have less priority.</i></p> <p><b>Environmental Impact:</b> Choices regarding the adoption of greener technologies, such as electrification of tracks or the use of energy-efficient trains, can significantly reduce the carbon footprint of railway operations over the long term.</p> <p><b>Financial Health:</b> Strategic decisions, whether related to operational efficiencies, expansions, or service offerings, can impact the financial health of railway operators, affecting their ability to invest in future improvements and innovations.</p> <p><b>Regulatory Compliance:</b> Ensuring compliance with current and anticipated regulatory requirements can mitigate the risk of future legal and financial penalties while also enhancing safety and operational standards.</p> <p><b>Workforce Development:</b> Investments in training and development, along with the adoption of new technologies, can enhance the skills and adaptability of the railway</p>

Air Traffic	Electricity	Railway
		<p>workforce, impacting the quality of operations and innovation capacity.</p> <p><b>Technological Advancement:</b> Decisions to implement advanced technologies, such as AI and IoT, can set a foundation for continuous innovation, transforming operations, maintenance, and customer service in lasting ways.</p> <p><b>Public Perception:</b> The way railway operators handle safety, environmental concerns, and customer service can influence public perception and trust in the railway system, affecting ridership and public support over the long term.</p> <p><b>Market Position and Competitiveness:</b> Strategic decisions can affect the competitive position of railway operators within the transportation market, influencing their ability to attract passengers and freight customers in competition with other modes of transport.</p> <p><b>Adaptability to Future Challenges:</b> Decisions that incorporate flexibility and scalability can prepare railway systems to adapt more effectively to future challenges, including</p>

Air Traffic	Electricity	Railway
		technological changes, shifts in demand, and regulatory developments.

## EVALUATION OF A DECISION

All decisions made regarding a certain context are evaluated ex-post according to certain criteria. In the context of the project, criteria are defined by several KPIs in the use cases, that are listed hereafter.

Category	Air Traffic	Electricity	Railway
Technical quality of AI-based solutions	Reduction in Delays	Assistant alert accuracy ( <i>or Assistant self-awareness</i> ) Operation score Network utilization Action recommendation selectivity Carbon intensity Topological action complexity	Response time (UC.01) Punctuality (UC.01) Delay Reduction Efficiency
Quality of AI-based solutions perceived by human operators	Prompt demand rate Significance of human revisions Efficiency score Acceptance score Agreement score Decision Support satisfaction	Assistant relevance Situation awareness	Acceptance score Acceptance Assistant relevance Comprehensibility Situation awareness
Human-AI interaction	AI co-learning capability	Human control and autonomy over the process	Human control and autonomy over the process

Category	Air Traffic	Electricity	Railway
		Human learning Decision support for the human operator	Human learning Decision support for the human operator
The efficiency of combined human-AI performance		Total decision time Ability to anticipate	Ability to anticipate Response time (UC.02) Punctuality (UC.02)
Cognitive load	Workload perception Human Response Time	Workload Assistant disturbance	Human Information Processing
Robustness	AI prediction robustness	Technical robustness to real-world imperfections Resilience to real-world imperfections Transferability across fidelity levels Generalization to different grids	
Trustworthiness	System Reliability Trust in AI solutions score	Trust towards the AI Tool Human motivation	Trust towards the AI Tool Human motivation

## ANALYSIS OF DECISION-MAKING SCENARIO

To identify common steps of the decision-making process across all domains, examples of decision-making scenarios have been described and analyzed for each domain: the aim is to illustrate the decision by breaking it down into several steps (as depicted in Figure 41).

This does not correspond to the use cases that have been defined, but to other scenarios that illustrate at best how decisions are made. The analysis of these scenarios has been made during workshops where the different elements have been created as post-its and then clustered on a common story map (see Figure 42). For each relevant scenario, a short description is provided, and only the relevant decision steps. Then, for each decision step, its characteristics are provided.

The result of this work gave birth to 5 clusters, which can be classified temporally:

1. Context (environment)
2. Event and trigger (that necessitate a decision)
3. Decision exploration
4. Decision validation / Feedback
5. Impact and evaluation (of the decision)

These clusters have led to the definition of the sub-sections of Section 3.1.2, and helped framing the 5 steps pattern in the abstract base user story of Section 3.2.3.1.6 : context, trigger and three actions.

Besides, this work has also allowed the identification of interesting components of human-computer interactions (HCI) and tasks carried out by the AI assistant. The story map thus displays elements from both:

- Human operator point of view (scenarios elements per domain, HCI),
- AI assistant point of view.

In addition to the analysis of the common human operator – AI assistant decision-making process, interesting elements of the AI decision exploration steps have been identified, such as the following:

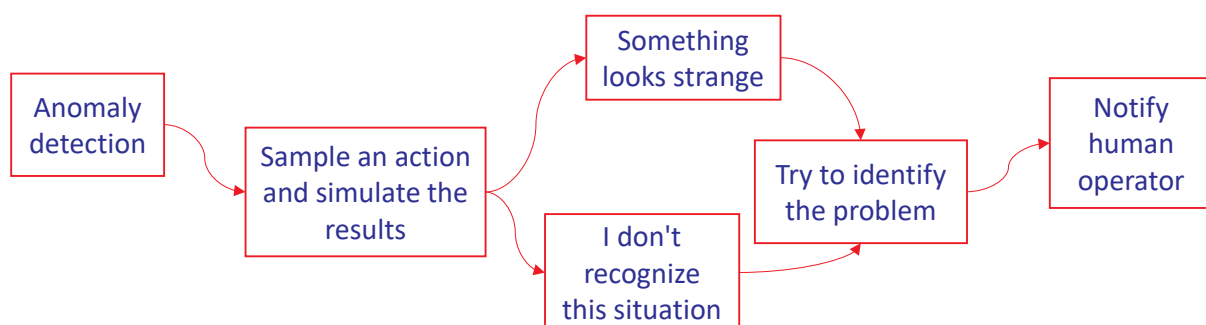


FIGURE 41 - AI DECISION EXPLORATION STEPS EXAMPLE

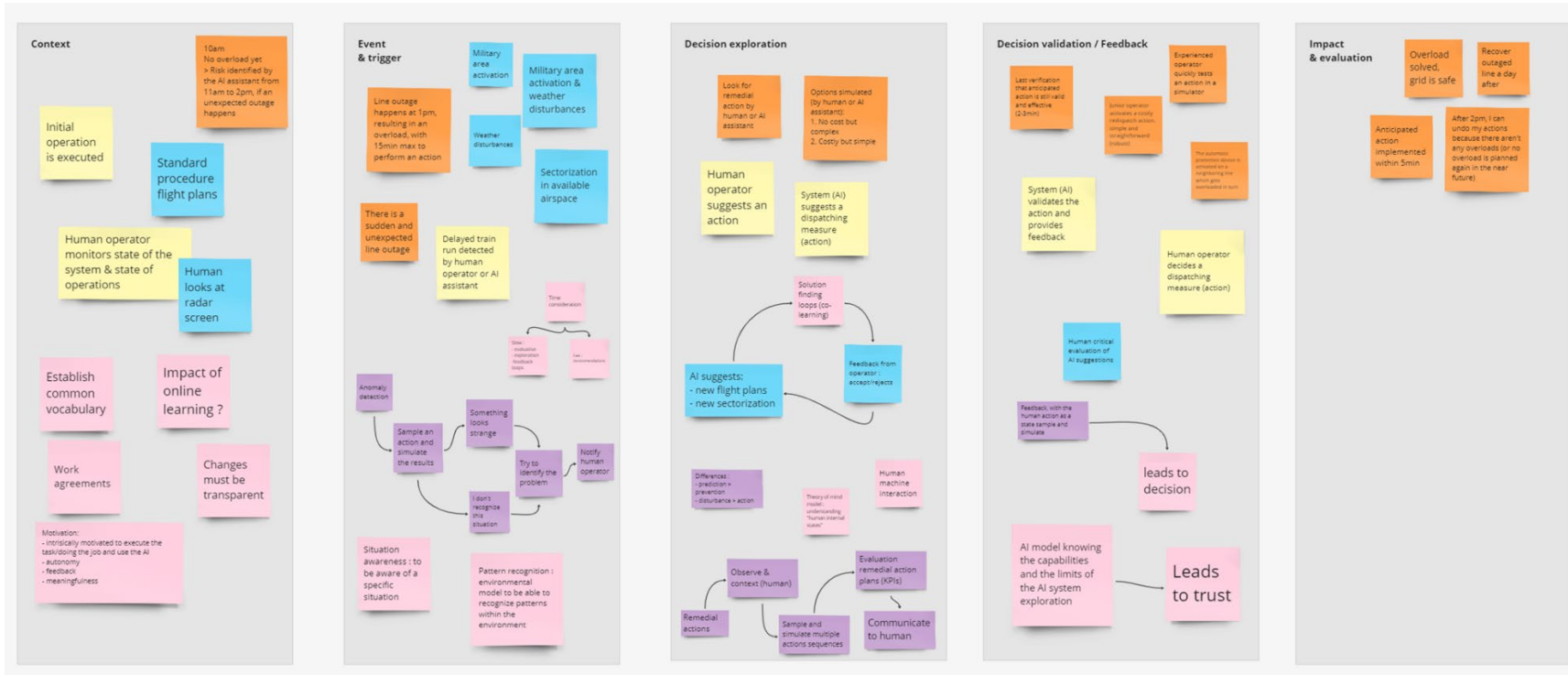
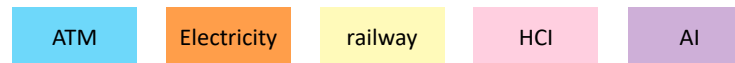


FIGURE 42 - COMMON ANALYSIS OF DECISION-MAKING SCENARIOS

The different elements on the board are identified by the following colors:





## AIR TRAFFIC

**Short description of the scenario:** the scenario is applicable to both use cases

### List of decision steps for ATM:

Step	Description	Type	Tradeoffs	Time constrains	Implementation	Action type
<b>Shift start</b>	Initial sectorization plan	-	-	-	-	-
<b>Real-time monitoring</b>	Monitoring airspace	Tactical/pre-tactical/operational	-	-	-	-
<b>New constraint#1 - Adverse weather report</b>	Readjust operational steps	Tactical/pre-tactical/operational	Re-route Airspace conditional management	Real time	Real time	Corrective/adaptative
<b>Process constraint#1</b>	Sectorization Routing	operational	Overload sectors Lower capacity	Real-time	Real-time	Corrective/adaptative
<b>New constraint#2 – capacity overload risk</b>	Capacity exceeds limits	Tactical/pre-tactical (2h)	Overload sectors	Pre-tactical (2h) to Tactical (real-time)	Before Capacity overload	Adaptative
<b>Process constraint#2</b>	Sectorization adaptation Flight re-routing	Operational	Capacity vs safety	-	Real-time	Adaptative
<b>Return to nominal conditions</b>	Normal operations	-	-	-	-	-

## POWER GRID

### Short description of the scenario: handling a new forecasted overload

Shortly after the beginning of his/her shift in the morning, the operator is made aware that a potential overload could occur on a transmission line.

#### References:

- [L2RPN challenge, Paris Region AI Challenge for Energy Transition](#), April 2023 (chapter §4)
- Work Domain Analysis of Electric Transmission, Networks and Operation, A. Hilliard, R. Brath, and G. A. Jamieson <https://doi.org/10.1109/JSYST.2023.3339709>

#### List of decision steps:

*The scenario hereafter is illustrated with a fictitious example of business operation context*

Step	Description	Type	Tradeoffs	Time constrains	Implementation	Action type
<b>Start (08:00AM)</b>	<p>Beginning of shift: previous operator has ended his/her shift.</p> <p>Planned outage beginning at 09.00 requires 2 actions.</p> <p><b>P1:</b> Change topology in an adjacent substation</p> <p><b>P2:</b> Coordinate and validate a transit limitation with a DSO</p> <p>Opportunity to improve voltage plan (decrease losses)</p>	-	-	-	-	-
<b>New alert forecasted at 10:00AM</b>	<p>A potential overload could occur starting at 10:00 on the line L1. This overload, if confirmed, needs a remedial action (else operational limits would be violated)</p> <p>Multiple solutions exist.</p>	-	-	-	-	-

Step	Description	Type	Tradeoffs	Time constrains	Implementation	Action type
<b>Processing the new alert (1)</b>	<p>Different remedial actions are possible</p> <p><b>R1:</b> load transfer from DSO (LTTD @08:15)</p> <p><b>R2:</b> change of topology in substation S1 (LTTD ~@09:40)</p> <p><b>R3:</b> limitation of RES generation (costly, LTTD ~@09:50)</p> <p>R2 seems the best option</p> <p>The operator decides to ignore R1 and wait</p>	Tactical	<p>Compliance with the operating limits of the components of the electrical system (reconfiguring the grid, which can wear out components) vs maintaining the operational safety of the power grid (overload)</p> <p>Long term and uncertain events (overload) vs complex actions (remedial actions)</p>	Next LTTD in more than 1 hour	Planned	Preventive
<b>Preparing the planned outage (1)</b>	<p>The operator prepares action P1 for the planned outage:</p> <ul style="list-style-type: none"> <li>Simulation of flows with changed topology</li> <li>Action list to change the topology</li> </ul>	Operational	<p>Guaranteeing the completion of maintenance work (planned outage) vs maintaining the operational safety of the power grid (forecasted overload)</p> <p>Short term and certain events (planned outage) vs longer term and uncertain events (overload)</p>	Planned outage in less than 1 hour	Planned	Preventive
<b>Preparing the planned outage (2)</b>	<p>The operator prepares action P2 for the planned outage:</p> <ul style="list-style-type: none"> <li>Topology with simulation of agreed load transfer from DSO</li> <li>DSO contact information</li> </ul>	Operational	(see previous step)	Planned outage in less than 1 hour	Planned	Preventive
<b>Processing the new alert (2)</b>	<p>The operator is evaluating another remedial action (R4) in the simulation tool: Load transfer, LTTD @09:30, which is more complex than R2, R3</p>	Tactical	Simple actions vs complex actions	Next LTTD in less than 1 hour	Planned	Preventive

Step	Description	Type	Tradeoffs	Time constrains	Implementation	Action type
<b>Processing the new alert (3)</b>	The operator decides to ignore R4's LTTD	Tactical	Simple actions vs complex actions		Planned	Preventive
<b>Alternative end #1</b>	At 09:45, overload is still forecasted  Given the short time remaining and the simplicity of R2, the operator decides to perform R2	Tactical	Compliance with the operating limits of the components of the electrical system (reconfiguring the grid, which can wear out components) vs maintaining the operational safety of the power grid (overload)  Given the short timeframe, there is a high probability that the overload happens: wait and decide later and rely on only one available remedial action vs act know and rely on more remedial actions.		Planned	Preventive
<b>Alternative end #2</b>	At 09:45, overload is not forecasted anymore	-	-	-	-	-

## RAILWAY

### Short description of the scenario: AI-driven Timetable Creation and Real-time Adjustment for Urban Rail Network

A rail operator implements an AI-based system to autonomously create and adjust train schedules in real-time, enhancing efficiency and punctuality across the network. The AI system is designed to manage the entire timetable, optimizing for peak and off-peak flows, and dynamically responding to delays, equipment failures, or sudden changes in passenger demand. Human intervention is reserved for major incidents or complex situations beyond the AI's decision-making capabilities.

#### List of decision steps:

Step	Description	Type	Tradeoffs	Time constraints	Implementation	Action type
<b>Initial Timetable Creation</b>	The AI analyzes historical data and the booked trips to create an optimal timetable.	Strategic	Efficiency vs. passenger needs	Weeks to months	Planned	Preventive
<b>Real-time Monitoring</b>	Continuous monitoring of the network through sensors and data sources.	Operational	Real-time accuracy vs. data overload	Real-time	Real-time	Preventive
<b>Delay Detection</b>	The AI detects delays and analyzes their impact.	Operational	Speed of response vs. accuracy of impact analysis	Immediate	Real-time	Corrective
<b>Dynamic Timetable Adjustments</b>	The AI recalculates the timetable to minimize delay impacts.	Tactical	Optimizing network efficiency vs. minimizing passenger inconvenience	Minutes to hours	Real-time	Corrective
<b>Communication and Implementation</b>	Automated alerts and updated schedules are communicated to passengers and staff.	Operational	Clarity and reach of communication vs. immediacy	Immediate to short-term	Real-time	Corrective

Step	Description	Type	Tradeoffs	Time constrains	Implementation	Action type
<b>Major Incident Escalation</b>	The AI escalates major incidents to human operators with data and analysis.	Strategic/Operational	AI decision-making capacity vs. complexity of human judgment	As needed	Real-time to planned	Corrective
<b>Post-Incident Analysis and Learning</b>	The AI analyzes responses to improve future performance.	Strategic	Learning accuracy vs. operational continuity	Post-incident analysis	Planned	Preventive