Toward a Holistic Framework for Human-AI Collaboration in Safety-Critical Systems

Ricardo J. Bessa, Milad Leyli-Abadi, Mouadh Yagoubi, Daniel Boos, Clark Borst, Alberto Castagna, Ricardo Chavarriaga, Duarte Dias, Adrian Egli, Andrina Eisenegger, Joost Ellerbroek, Anna Fedorova, Cristina Felix, Anton Fuxjäger, Joaquim Geraldes, Samira Hamouche, Mohamed Hassouna, Sjoerd Kop, Bruno Lemetayer, Giulia Leto, Roman Liessner, Jonas Lundberg, Antoine Marot, Maroua Meddeb, Manuel Meyer, Hélio Sales, Viola Schiaffonati, Manuel Schneider, Irene Sturm, Julia Usher, Herke Van Hoof, Jan Viebahn, Toni Wäfler, and Giacomo Zanotti

Abstract The integration of artificial intelligence (AI) into safety-critical systems, where human operators remain central to decision-making, introduces various challenges that existing AI frameworks struggle to address comprehensively. Key concerns involve designing a socio-technical system that balances AI transparency, trust, and explainability with the imperative for robust and reliable decision-making.

R. J. Bessa (✉) · D. Dias
INESC TEC, Porto, Portugal
e-mail: ricardo.j.bessa@inesctec.pt; duarte.f.dias@inesctec.pt

M. Leyli-Abadi · M. Yagoubi · M. Meddeb
IRT SystemX, Paris, France
e-mail: milad.leyli-abadi@irt-systemx.fr; mouadh.yagoubi@irt-systemx.fr; maroua.meddeb@irt-systemx.fr

D. Boos · A. Egli
SBB Swiss Federal Railways, Bern, Switzerland
e-mail: daniel.boos@sbb.ch; adrian.egli@sbb.ch

C. Borst · J. Ellerbroek · G. Leto
Delft University of Technology, Delft, The Netherlands
e-mail: c.borst@tudelft.nl; j.ellerbroek@tudelft.nl; G.Leto@tudelft.nl

A. Castagna · A. Fuxjäger
EnliteAI, Wien, Austria
e-mail: a.castagna@enlite.ai; a.fuxjaeger@enlite.ai

R. Chavarriaga · A. Fedorova
Zurich University of Applied Sciences, Zurich, Switzerland
e-mail: char@zhaw.ch; fedo@zhaw.ch

E. Curry et al. (eds.), *Artificial Intelligence, Data and Robotics*,
https://doi.org/10.1007/978-3-032-10561-5_13

A. Eisenegger · S. Hamouche · J. Usher · T. Wäfler
University of Applied Sciences and Arts Northwestern Switzerland, Olten, Switzerland
e-mail: andrina.eisenegger@fhnw.ch; samira.hamouche@fhnw.ch; julia.usher@fhnw.ch; toni.waefler@fhnw.ch

C. Felix · J. Geraldes · H. Sales
NAV Portugal, Lisbon, Portugal
e-mail: Cristina.Felix@nav.pt; Joaquim.Geraldes@nav.pt; Helio.Sales@nav.pt

M. Hassouna
Fraunhofer IEE, Kassel, Germany
e-mail: mohamed.hassouna@iee.fraunhofer.de

S. Kop · J. Viebahn
TenneT, Arnhem, The Netherlands
e-mail: sjoerd.kop@tennet.eu; jan.viebahn@tennet.eu

B. Lemetayer · A. Marot
Réseau de Transport d'Electricité, Paris, France
e-mail: bruno.lemetayer@rte-france.com; antoine.marot@rte-france.com

R. Liessner · I. Sturm
Deutsche Bahn, Berlin, Germany
e-mail: roman.liessner@deutschebahn.com; Irene.Sturm@deutschebahn.com

J. Lundberg
Linköping University, Norrköping, Sweden
e-mail: jonas.lundberg@liu.se

M. Meyer · M. Schneider
Flatland Association, Bern, Switzerland
e-mail: manuel.meyer@flatland-association.org; manuel.schneider@flatland-association.org

V. Schiaffonati · G. Zanotti
Politecnico di Milano, Milan, Italy
e-mail: viola.schiaffonati@polimi.it; giacomo.zanotti@polimi.it

H. Van Hoof
University of Amsterdam, Amsterdam, The Netherlands
e-mail: h.c.vanhoof@uva.nl

Presently, while numerous sector-specific solutions exist, a holistic framework that effectively integrates human expertise with AI capabilities remains absent, leaving critical gaps in system design, deployment, and oversight. This chapter proposes a multidisciplinary conceptual framework to enhance human-AI collaboration in critical infrastructures such as power grids, railways, and air traffic management. The different design steps were guided by the requirements of these industrial domains. The framework combines key design principles that support human cognition, leveraging insights from decision theory, mathematics, and specialized engineering domains to optimize AI-assisted decision-making. Furthermore, it embeds trustworthiness and risk assessment methodologies, using tools such as the Assessment List for Trustworthy Artificial Intelligence (ALTAI) tool to ensure compliance with ethical and regulatory requirements.

Keywords Safety-critical systems · Framework · Human centric · Power grid · Railway · Air traffic management

Acronyms

AI	Artificial intelligence
ALTAI	The Assessment List for Trustworthy Artificial Intelligence
ATCOs	Air traffic controllers
ATM	Air traffic management
CAB	Cockpit and bidirectional assistant
CSE	Cognitive systems engineering
EID	Ecological interface design
HMI	Human-machine interface
JCF	Joint Control Framework
KPIs	Key performance indicators
MBSE	Model-based system engineering
RL	Reinforcement learning
TAI	Trustworthy AI
UQ	Uncertainty quantification
UTM	Unmanned aircraft system traffic management
XAI	Explainable AI

1 Introduction

In the context of emerging climate change impacts and growing demand for energy and mobility networks, complemented with the implementation of digitalization and decarbonization roadmaps, artificial intelligence (AI) can be a transformative technology in the management and operation of critical infrastructures. These sectors are classified as high risk in the European Union AI Act. Therefore, safety, transparency, and adherence to ethical principles are fundamental to mitigating risks across the technical, social, and environmental dimensions. Since humans play a pivotal role in operational decision-making within these infrastructures (e.g., train redispatch, power grid congestion management), achieving seamless human-AI collaboration remains a critical challenge. This requires integrating human expertise into AI-driven processes, promoting mutual adaptation and shared decision-making while preventing individual decision-making. Various methodologies can be employed, including co-learning paradigms where AI and human operators iteratively enhance their knowledge and hybrid systems that enable AI to function autonomously while maintaining human oversight for critical decisions.

This chapter describes a conceptual framework tailored for the operation of critical infrastructures aided by AI-based systems and relates to the *cross-sectional* *reasoning and decision-making* areas of the AI, Data, and Robotics Partnership [1]. The framework incorporates AI algorithms, AI-friendly digital environments, a hypervision-based human-AI interface, and socio-technical design principles to enhance seamless human-machine collaboration. It aims to enhance real-time and predictive operational efficiency while offering structured methodologies for managing uncertainty and mitigating operational risks. Operators of critical infrastructures and academia designed this framework for the electrical power grid, railway network, and air traffic management.

1.1 AI Frameworks for Power Grid Operations

Marot et al. introduced a framework for AI-driven agents to assist in real time human operators in solving technical problems, such as power line overloads in power grids. There, the AI agent proactively alerts the operator when the confidence level in its suggested remedial actions is low, thereby preventing a human-out-of-the-loop scenario [2]. To reduce excessive alarms, an attention budget mechanism was implemented. The framework combines three key principles: *credibility*, achieved via transparency and clear explanations of AI-driven actions; *reliability*, maintained via consistent performance and generalization across different scenarios; and *intimacy*, where detailed explanations for incorrect actions are provided to build trust between human operators and AI. A framework based on naturalistic decision-making is presented in Greitzer and Podmore [3], combining concepts such as recognition-primed decision-making, recognition/meta-recognition, and situational awareness. It models how power grid operators process information, interact with environmental cues, and make decisions in real time. Two internal simulation loops allow for rapid meta-cognitive evaluation and prediction of the potential effects of control actions, thereby supporting decision-making in grid operations. Fan et al. introduce a human-machine hybrid-augmented intelligence framework for grid dispatching and control [4]. It integrates AI-driven decision aid, a digital twin, and human expertise to enhance operational efficiency in dynamic and complex grid management tasks. Key elements include bidirectional data exchange, knowledge co-evolution between AI and humans, and flexible task redistribution between operators and AI systems to optimize decision-making. Hilliard et al. propose the Work Domain Analysis framework that structures power grid operations into five hierarchical levels, linking physical grid components to overarching objectives such as reliability and safety [5]. The framework employs means-ends relationships, hierarchical decomposition, and cause-effect modeling to enhance operator decision support, improve the visual representation of grid states, and facilitate structured decision-making.

1.2 AI for Railway Traffic Management

Traditionally, traffic management in railway networks is performed by human operators, nowadays supported by traffic management systems (TMS) with various degrees of automation, such as the Swiss Federal Railway's Rail Control System [6], which provides real-time traffic monitoring and dynamic routing options to the human dispatcher. While such TMS leverage the long history of operations research in the railway domain, experimental systems like the automatic dispatching assistant developed by Deutsche Bahn [7, 8] use machine learning to enhance the human decision support. In recent years, there has been growing interest in machine learning approaches, such as deep reinforcement learning (DRL) and multi-agent reinforcement learning (MARL) in particular, for planning and (re-)scheduling tasks in railway [9], and fundamental features of DRL for vehicle rescheduling problems [10] in railway were demonstrated. For example, the international community of researchers and railway enthusiasts around the Flatland digital environment [11] illustrated basic generalization and scalability capabilities of DRL through open machine learning challenges [12]. In addition, researchers showed the effectiveness of MARL for real-time rescheduling [13], how advanced machine learning model architectures enhance coordination between trains [14], and how communication between MARL agents can improve solution quality [15]. Researchers further demonstrated that MARL-based rescheduling can save energy in train operations [16]. While DRL-based traffic management remains an active area of research, a recent study showcased the application of MARL for real-world scenarios [17], and novel open-source tools such as the Open Source Railway Designer [18] support the translation of basic research results to real-world settings.

1.3 Framework for Air Traffic Management (ATM) Systems

Currently, ATM remains predominantly human centric, with responsibility for operational safety resting on human operators, including strategic flight planners and tactical Air Traffic Controllers (ATCOs). Nunes et al. [19] developed an AI-based support system tailored for tactical ATCOs, aiming to align conflict detection and resolution recommendations (CD&R) with human decision-making strategies and expectations. The system uses images as inputs for a Reinforcement Learning (RL) agent, allowing it to generate a human-interpretable state representation for learning. Project MAHALO [20] studied the role of transparent and conformal AI in CD&R, showing that automation designed to mimic human strategies closely—referred to as conformal automation—is generally more accepted than purely transparent AI. Despite the benefits of such tools in reducing ATCOs workload, human-in-the-loop experiments have highlighted challenges such as decreased situational awareness and issues related to trust and acceptance. Notably, greater transparency did not always resolve these concerns, as an increased flow of information sometimes resulted in delayed responses, which is particularly problematic in time-sensitive operations. In Unmanned Aircraft System Traffic Management (UTM), where automation plays a dominant role due to the absence of human pilots, AI-driven multi-agent path planning has demonstrated significant potential in optimizing flight routes and resolving conflicts [21, 22]. Additionally, some studies have explored implementing adaptive sector structures, aiming to enhance workload distribution among controllers using heuristic and learning-based optimization techniques [23, 24].

1.4 Socio-Technical Systems and Human-AI Interaction

Frameworks addressing the socio-technical dimensions of decision-making are becoming increasingly relevant. The Joint Control Framework (JCF), for example, emphasizes collaborative decision-making between humans and AI systems [25]. It provides methodologies for adaptive control and co-learning, ensuring that AI-driven decisions incorporate human expertise, especially in high-stakes situations where full automation is impractical.

1.5 Trustworthy AI Frameworks

Trust plays a fundamental role in decision-making frameworks for critical infrastructures. The Confiance.AI framework [26, 27] and the Human AI Ethical Framework [28] highlight the necessity of developing AI systems that uphold transparency, ethical standards, and human values. These frameworks establish principles to ensure that AI-driven decision-making systems remain reliable, interpretable, and aligned with societal expectations.

1.6 Risk Management and Regulations

The Assessment List for Trustworthy Artificial Intelligence (ALTAI) tool, the European Union's AI Act, and ISO/IEC 42001 AI management provide a structured methodology for evaluating and mitigating the risks associated with AI deployment in critical infrastructures [29]. These regulatory frameworks facilitate the assessment of both ethical considerations and technical functionalities, ensuring that AI-driven decision processes adhere to stringent safety, accountability, and compliance standards.

All these frameworks offer valuable and different perspectives on integrating AI into decision-making processes within critical infrastructures. However, their domain-specific focus often results in fragmented approaches, addressing isolated challenges rather than providing a unified solution. There is an increasing demand for a comprehensive conceptual framework that accommodates the diverse requirements of critical infrastructures, promoting a holistic decision-making strategy that integrates technical, ethical, and human-centered considerations. The proposed framework results from the European project AI4REALNET (*AI for REAL-world NETwork operation*) and focuses on developing AI-driven decision systems that are not only technically sound but also ethically aligned and context aware, improving both social and technical dimensions of system performance. Furthermore, it covers three distinct levels of interaction between humans and AI: (a) full human control with AI assistance, (b) co-learning where AI and human operators continuously refine their collaboration, and (c) autonomous AI with human supervision.

The design and implementation of this framework are guided by industry-relevant use cases that address key operational challenges network operators face. These use cases, described in the Appendix section, are grounded in real-world operational settings where human oversight plays a crucial role. This framework is intended for use by various stakeholders, including AI developers, innovation managers, network operation managers, regulatory authorities, and standardization organizations.

An interdisciplinary approach is adopted to develop the proposed holistic conceptual framework for decision-making in critical infrastructures by integrating traditionally distinct fields such as psychology and cognitive engineering. The framework also drew on mathematics, decision theory, computer science, and specialized engineering domains, particularly energy and mobility. A high-level overview of the proposed conceptual framework is illustrated in Fig. 1, which also outlines the structure of this chapter.

The systems engineering and theories are adapted for trustworthy AI integration in designing the conceptual framework's operational, functional, and logical

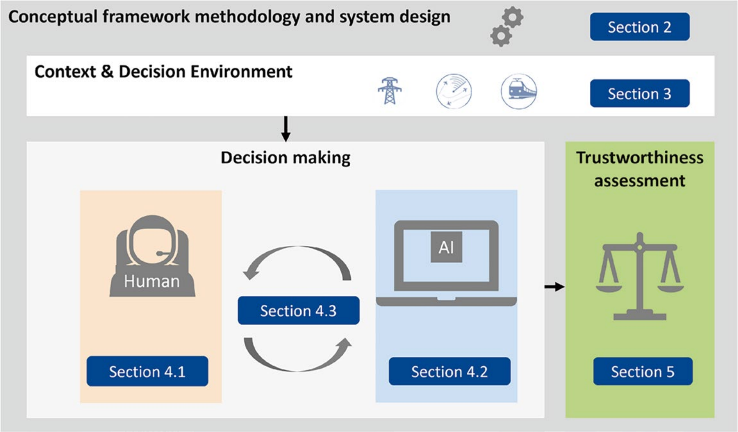

Conceptual framework methodology and system design
Section 2
Context & Decision Environment
Section 3
Decision making
Human
Section 4.1
Section 4.3
AI
Section 4.2
Trustworthiness assessment
Section 5

Fig. 1 The proposed conceptual framework, comprising three main components governed by system design principles and their corresponding sections

architectures to meet both functional and non-functional requirements of critical infrastructures (Sect. 2). Based on the observed context and decision environment in a specific industrial domain (Sect. 3), the decision-making process (Sect. 4) is responsible for ensuring the stability and resilience in the corresponding critical infrastructure when an event is triggered. Decisions may be the result of collaboration between human operators and AI-based assistants through a provided interface. As such, the human cognitive process could be increased by AI-based assistants through different perspectives. Furthermore, the decisions must comply with established trustworthy key performance indicators (KPIs) and undergo validation by a regulatory authority (Sect. 5). The subsequent sections provide a detailed explanation of each component of the framework.

2 Conceptual Framework Methodology 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. Recent advances in artificial intelligence enable the development of systems capable of performing increasingly complex tasks, opening new opportunities to assist industries in various sectors while raising important ethical and social issues; the design of these systems is crucial to ensure their effectiveness and acceptance [30, 31]. To cope with the complex environment and tasks the system will operate in and execute, multiple viewpoints (of stakeholders) on the envisioned system are taken to derive both functional and non-functional requirements. It also guides the development process of the proposed framework's submodules and their interactions. 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 human-AI 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*.

These initial views, definitions, and considerations outlined in this section play a crucial role in ensuring the overall coherence and alignment of the system between the different aspects during development. Figure 2 summarizes the adopted

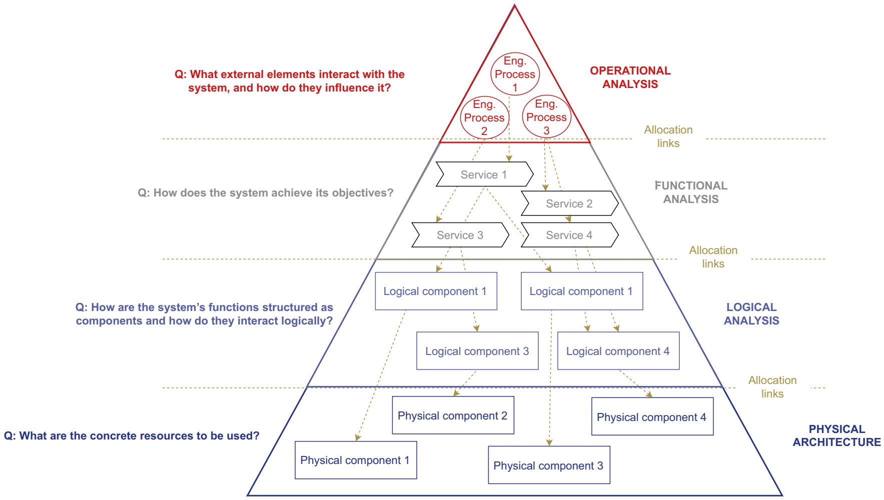


Fig. 2 The Arcadia modeling pyramid used for system design of the conceptual framework

Model-Based System Engineering (MBSE) approach using the ARCADIA method [32]. With the aim of proposing a conceptual framework, the physical architecture at the lowest level of the pyramid is not considered. Each of the operational, functional, and logical analyses with corresponding design processes is described in the following.

2.1 Operational View

The operational view focuses on how the system functions in real-world scenarios. It addresses the operational aspects of the system, including the interactions between users, systems, and external entities, and uses various diagrams to visually represent them. It also allows us to identify the operational, functional, and non-functional requirements.

Using this analysis, various stakeholders involved in the three previously mentioned critical infrastructures are primarily identified. These stakeholders may interact with the proposed frameworks during the different phases of the decision-making process and could be attributed to different roles. For example, an operator could be responsible for implementing a decision and also providing the AI-based system with some feedback at the same time. In the following, these stakeholders and their interactions with the conceptual framework are analyzed. It is followed by the identification of operational requirements and those based on cooperation with AI-based assistance, which may increase productivity and human capacity.

2.1.1 Operational Environment Diagram

An operational environment diagram in system design visually represents the external entities (e.g., stakeholders), interactions, and contexts in which a system operates, helping define its boundaries and interfaces. Figure 3 illustrates these interactions, depicting the system as a black box with data flows between various stakeholders, including human operators, AI-based decision systems, and regulatory agents. This diagram is designed based on the analysis performed on three critical infrastructures.

The *environment* provides real-world context and data, which are utilized by different actors—*data profiles*, for instance, are responsible for training AI decision systems supporting human operators. *Operators* and *supervisors* interact with both the *environment* and *simulators* via the conceptual framework interface, to assess or monitor the impact of their actions before execution in real-world settings. *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. Additionally, operators engage directly with the framework throughout the decision-support process, requesting AI assistance in different operational scenarios. To ensure security and compliance, *regulatory agents* analyze system decisions to verify adherence to established guidelines and *standards*.

2.1.2 Operational Requirements of Three Critical Infrastructures

Operational requirements define the conditions and performance criteria a system must meet to function effectively in its intended environment. These requirements encompass functionality, performance, reliability, security, and user interactions and are specified for each industrial domain and its corresponding scenarios. They are categorized into functional (F) vs. non-functional (N) and generic (G) vs. specific (S) requirements. Generic requirements are directly integrated into the functional view of the framework, while specific requirements are generalized to ensure broader applicability across use cases. Table 1 presents an extract of this classification, aiding in the structured design of the framework's functional architecture.

2.1.3 Human-in-the-Loop and Oversight Requirements

In addition to the operational requirements intrinsic to the domain presented earlier, a set of functional requirements has been identified to enable human-in-the-loop (HITL) decision-making under risk and uncertainty. Table 2 lists these requirements, providing a short description and a categorization based on the part of the system and the actor, when possible, that must adhere to these requirements. The identified actors are *Operator*, the human interacting with the system; *Agent*, the AI-powered side of the decision-making process; and *Environment*, either the true

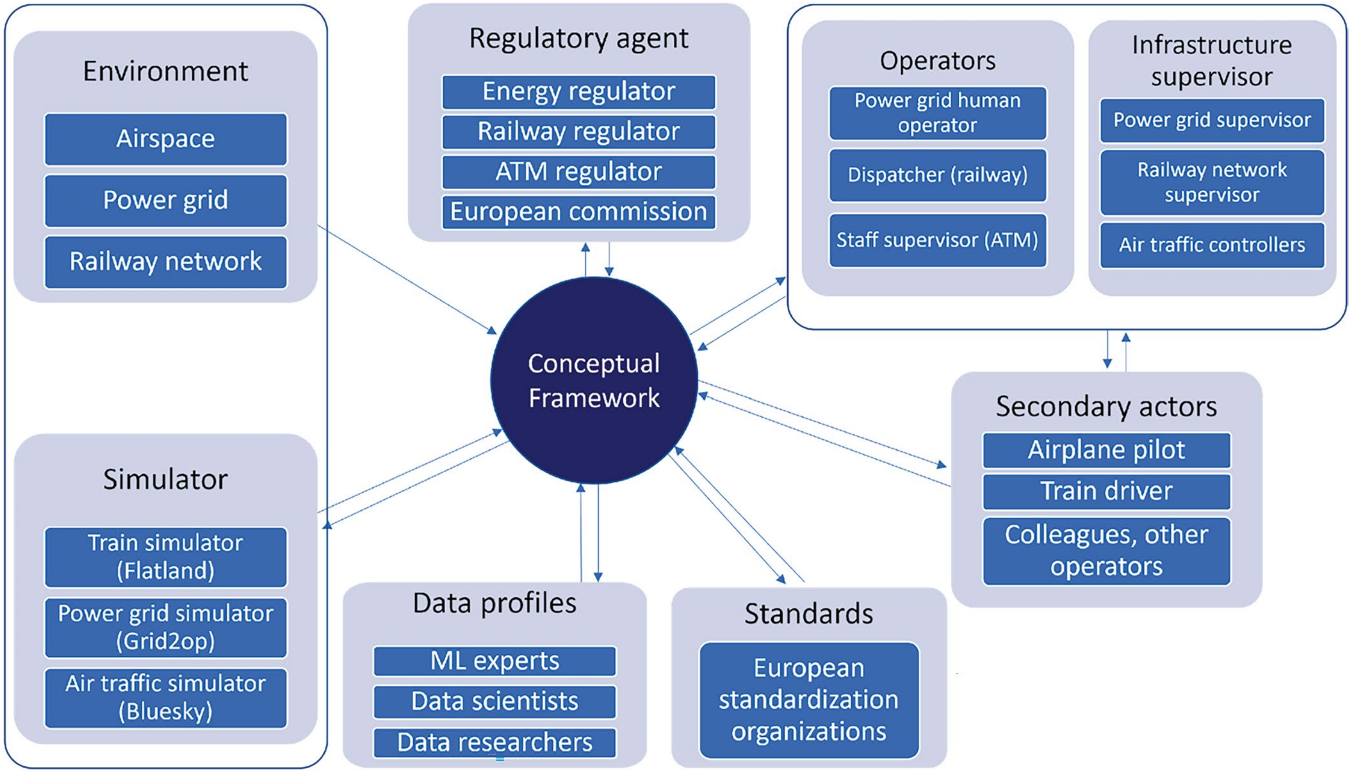


Fig. 3 Operational environment diagram illustrating the external entities, interactions, and contexts in which a system operates

Table 1 Operational requirements for three industrial domains

Category	Power grid	Railway	Air traffic
Robustness	• AI informs the human operator about its confidence in the output recommendation (self-awareness) (F, G) • Reproducibility and traceability of recommendations for *postmortem* analysis (F, G) • Robustness to attacks targeting model space and reward function (N, G)	• Reasonable recommendations in new situations (not seen during model training) (F, G) • Good performance in operating scenarios with high variability (N, G) • Retrospective quality control (N, G)	• System resilience to unexpected events (N, G) • Cyber and data security (N, G) • System's reliable operation and decisions (N, G)
Efficiency	• Computational efficiency (N, G) • Relevance of the recommendations (N, G) • Scalability (N, G) • Adequate training environment (N, G)	• Capacity to handle operating scenarios with high complexity (N, G) • Scalability (N, G) • Generalization to different scenarios (F, G)	• Capability to optimize resources and operations (F, S) • Scalability (N, G)
Interpretability	• Adaptability to different levels of interaction and human operator preferences (F, G) • Transparency during system training (F, G) • Capacity to explain recommendation(s) to the human operator (and other stakeholders) (F, G)	• Interpretability of suggestions (F, G)	• Provide clear, understandable explanations for its decisions (F, G) • Usability of the system from the human and other stakeholders' perspective (N, G)
Non-discrimination and fairness	• Avoid creating or reinforcing unfair bias in the AI system (F, G) • Regular monitoring of fairness (F, S)	• Distribution of delays (N, G)	×
Human agency and oversight	• Additional training about AI for human operators (N, G) • Mitigate addictive behavior from humans (N, G) • Mitigate de-skilling in the human operators (N, G)	×	×

Table 1 (continued)

Category	Power grid	Railway	Air traffic
Regulatory and legal	• Compliance with existing operational policies (N, G) • European AI Act (F, G) • Transparency to humans in terms of interaction with an AI system (N, G)	• Compliance with legal standards and regulations (N, G) • RUOM favoritism (N, S)	• Compliance with legal standards and regulations (N, G)
Data governance	• Processing of human operator data (N, G)	×	×
Accountability	• Allow audits for the AI recommendations and human operator actions (N, S) • Reporting of potential vulnerabilities, risks, or biases (F, G)	×	×
Other	×	• Maintainability (N, G) • Environmental sustainability (N, G)	• Maintainability (N, G) • Environmental sustainability (N, G)

one or the cloned copy accessed by the agent for forecasting and assessment. The identification of these requirements and their association with different actor classes facilitates the decision-making process in the context of the proposed interactive framework.

2.1.4 Abstract Base User Story

The user story allows us to clearly define a system's functionality from the end user's perspective, ensuring that development aligns with user needs and expectations. The user stories were derived during a cross-domain workshop to identify commonalities across analyzed critical infrastructures and guide the development of the AI-decision system, particularly in relation to the human-machine interface.

As can be seen in Fig. 4, the abstract base user story has three manifestations depending on the time horizon: *planning* (preparing for foreseen events), *near real time* (preventing predicted events), and *real time* (correcting events that have occurred). Each manifestation follows a structured sequence: within a given context, a trigger initiates three key actions. First, the situation is observed, either through real-time monitoring, simulation of potential futures, or identification of foreseen events. If a deviation or potential issue is detected, possible measures are explored. Finally, an intervention takes place, where the human operator, using the assistance provided by the AI-based system (depending on the considered

Table 2 Human-in-the-loop and oversight requirements

Categories ID	Category name for requirement	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 feedback	An operator can analyze the feedback provided by the autonomous agent to decide whether a recommendation should be followed	Operator
R-05	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-06	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-07	Simulate remedial plan	An agent can interact with a copy of an environment to simulate a remedial plan and/or provide feedback	Agent
R-08	Provide visual human readable state	An environment should provide a graphical depiction of the ongoing situation to allow human intervention	Environment
R-09	Estimate aleatoric uncertainty	An environment should support the estimation of aleatoric uncertainty derived from an external source, such as weather conditions	Environment
R-10	Raise alarm	The system should have an alarm accessible by a different range of actors to enable a human to intervene	Environment
R-11	Communicate alarm context	On the triggering, the system should provide extensive information about the causes that triggered the alarm	Other

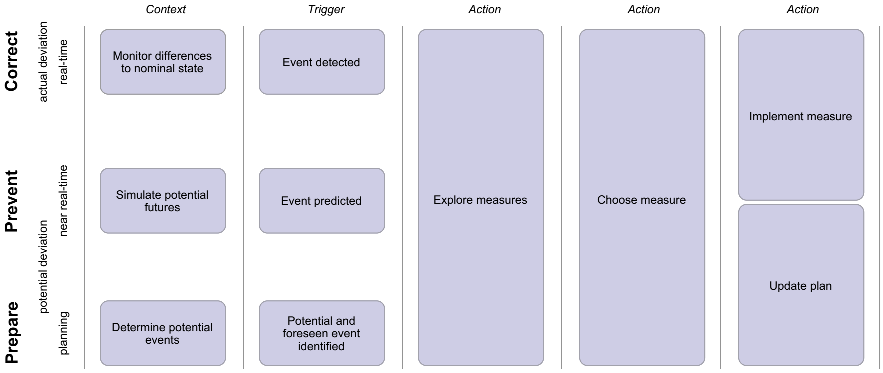


Fig. 4 Abstract base user story

interaction mode), implements a corrective or preventive action or integrates the measure into the operational plan for future execution. This structured approach ensures adaptability across different domains.

2.2 *Functional View*

The functional view answers the following question: “How does the system achieve its objectives?” Thereby, it defines the system functions and shows how system components process inputs, produce outputs, and interact. Based on the functional requirements identified in the operational analysis, the functional view contributes to the definition of system functions that satisfy those requirements through a functional decomposition approach. Next, it establishes how the identified functions interact with each other and with stakeholders while also outlining the associated data flow.

2.2.1 Functional Decomposition Diagram

Based on the identified functional requirements, the functional view outlines eight key functions as shown in Fig. 5. Each of these functions is broken down into subfunctions to ensure effective management of critical infrastructures and is identified based on the three studied industrial domains. These core functions include *interacting with the operator* to facilitate communication and support, *determining real-time context* by analyzing internal and external data, *selecting human-AI interaction*

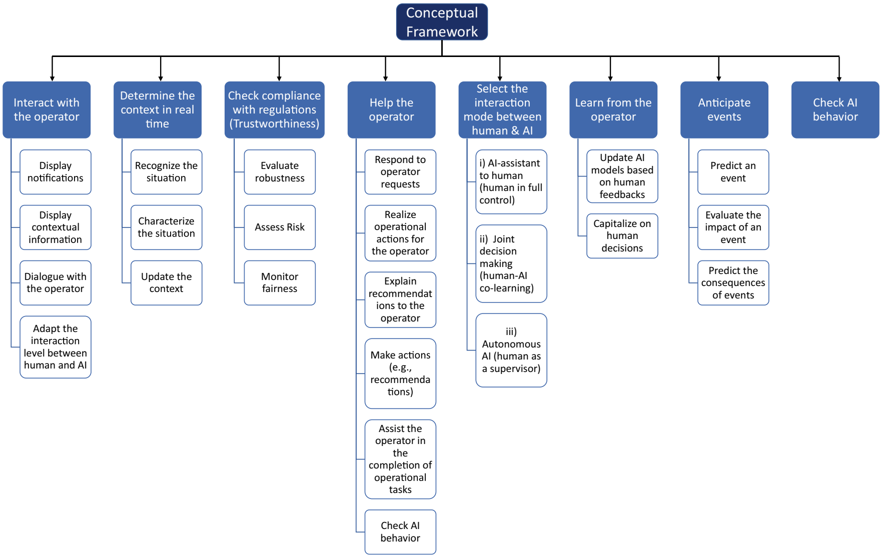

Conceptual Framework
Interact with the operator
Display notifications
Display contextual information
Dialogue with the operator
Adapt the interaction level between human and AI
Determine the context in real time
Recognize the situation
Characterize the situation
Update the context
Check compliance with regulations (Trustworthiness)
Evaluate robustness
Assess Risk
Monitor fairness
Help the operator
Respond to operator requests
Realize operational actions for the operator
Explain recommendations to the operator
Make actions (e.g., recommendations)
Assist the operator in the completion of operational tasks
Check AI behavior
Select the interaction mode between human & AI
i) AI-assistant to human (human in full control)
ii) Joint decision making (human-AI co-learning)
iii) Autonomous AI (human as a supervisor)
Learn from the operator
Update AI models based on human feedbacks
Capitalize on human decisions
Anticipate events
Predict an event
Evaluate the impact of an event
Predict the consequences of events
Check AI behavior

Fig. 5 Functional decomposition diagram

modes to customize collaboration, *assisting the operator* with decision-making, *learning from the operator*'s actions to enhance the system's knowledge base, *anticipating future events* to proactively address potential issues, ensuring *compliance with regulations* to maintain trustworthiness, and *continuously monitoring AI* behavior to detect and respond to anomalies. The next section illustrates the interaction and data flow between these functions and also with the external entities in the context of the human-AI decision-making process.

2.2.2 Functional Interaction Diagram

The functional interaction diagram (Fig. 6) illustrates the interaction and data flow between the identified system functions in the previous section, starting from high-level functions down to elementary ones that directly interact with stakeholders. It integrates the eight main functionalities while also incorporating interactions with the operator, simulator, and regulatory agent. As shown in the diagram, the operator can engage directly with the platform for assistance, select an interaction mode, and provide feedback on recommendations, which is then stored as new knowledge. The operator could simulate the impact of their actions using the simulators. When AI-based assistance is provided to increase the human cognitive process, their behavior is continuously monitored through a set of technical KPIs. A regulatory agent is also responsible for verifying the decision process and its compliance with standards and regulations.

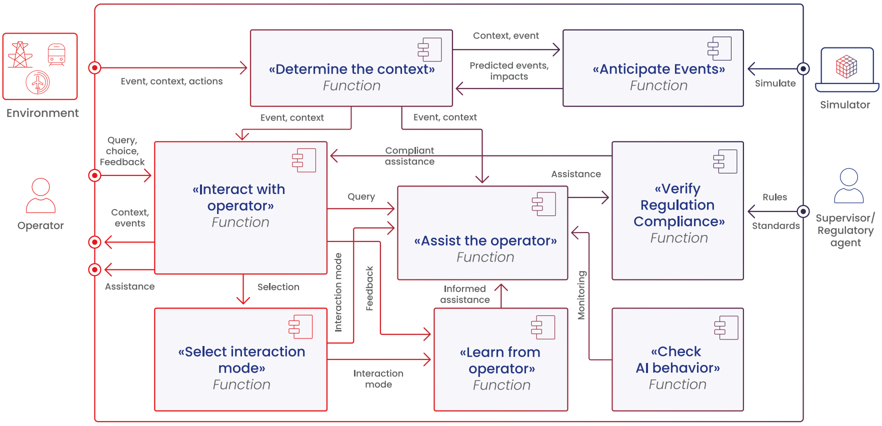

Environment
Operator
Simulator
Supervisor/ Regulatory agent
Event, context, actions
«Determine the context» Function
Context, event
Predicted events, impacts
«Anticipate Events» Function
Simulate
Event, context
Event, context
Query, choice, Feedback
«Interact with operator» Function
Compliant assistance
Assistance
«Verify Regulation Compliance» Function
Rules
Standards
Query
«Assist the operator» Function
Context, events
Assistance
Selection
Interaction mode
Feedback
Informed assistance
Monitoring
«Select interaction mode» Function
Interaction mode
«Learn from operator» Function
«Check AI behavior» Function

Fig. 6 Functional interaction diagram representing the relationships, dependencies, and information flow between system functions to illustrate how components interact to achieve overall system objectives and allow for determining the data flow among the eight identified key functions

2.3 Logical View

The logical view provides a blueprint for integrating the identified core functions, focusing on seamless interaction among subsystems (representing multiple functions performing a specific task) and stakeholders (internal or external). It also focuses on the abstract structure of the conceptual framework, defining the logical components and their interactions to meet system requirements. This view breaks down the system into three interaction modes: human in full control, joint human-AI decision-making (co-learning), and autonomous AI with human supervision. Each mode emphasizes varying levels of human involvement in decision-making processes, from full authority to collaborative learning and oversight of autonomous AI operations. This decomposition aids in understanding the system's functionality and supports stakeholders by presenting a clear, conceptual model. However, for the purpose of consistency, these three different interaction modes and their corresponding logical architecture are described in the decision-making section and under the human-AI interaction in Sect. 4.3.

In the following, the process view illustrating the generic architecture, including the subsystems and stakeholders, and their interactions, is presented. This generic process is then instantiated for each of the abovementioned interaction modes.

2.3.1 Process View

The generic process is presented in Fig. 7 as an overview of the high-level interaction between different subsystems that encompass one or more functions, identified during the functional analysis, to perform a task (e.g., AI-based agent including *assist the operator* and *learn from operator* functions). As shown in this process view, the *environment*, consisting of both the real-world setting and simulators, provides contextual information to the *human operator* and *AI-based* agent. The

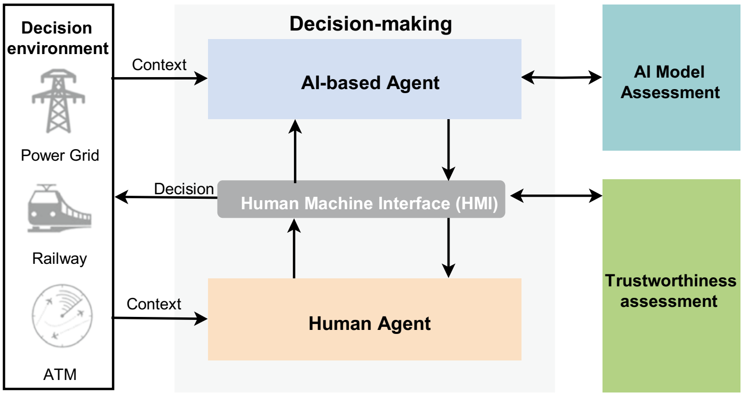


Fig. 7 Process view as an overview of the high-level interaction between different subsystems

AI-based agent assists the human operator via the *human-machine interface* (*HMI*) by offering recommendations or increasing his/her cognitive process, while its behavior is continuously monitored by the *AI Model Assessment* subsystem to ensure robustness and reliability. In return, the human operator can provide feedback, allowing the AI-based agent to refine its performance for future assistance. The interaction should also be conducted in a way to avoid the cognitive load of the human operator (see hypervision definition in Sect. 4.3.3). Finally, the human-AI decision-making process is evaluated through *trustworthiness assessment* KPIs and verified by a regulatory agent for regulatory compliance before being implemented in the decision environment.

2.3.2 Building Block View

As presented in the previous section, the conceptual framework is structured into various layers. 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, a high-level conceptual prototype is developed, which allows for testing and refining ideas, ensuring research outcomes meet practical needs. It aims to evolve and serve as initial design guidelines for future applications. This offers a hierarchical representation of the system from a technical perspective. Figure 8 shows the scope, context, and high-level view of the 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 submodule, a simulation engine, AI agents, and a digital environment. It receives current system data and requests simulations to predict events and consequences. The simulation results are evaluated and sent to the recommendation module, with all relevant data stored for future use.

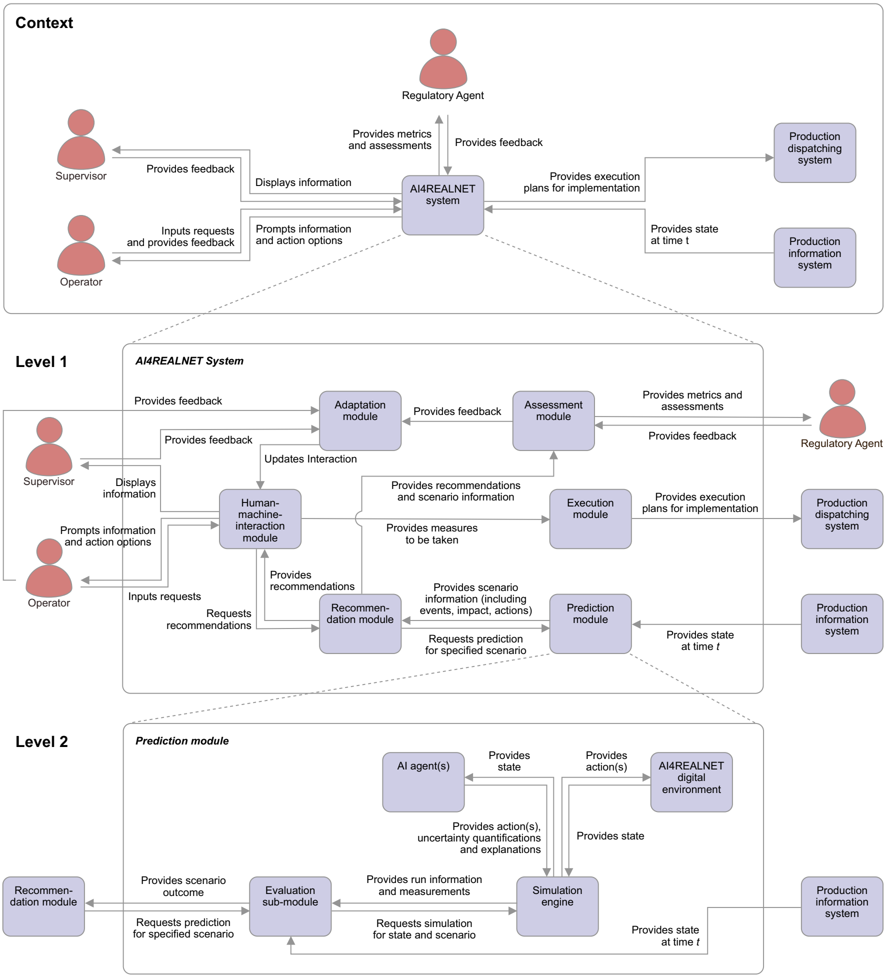

Context
Regulatory Agent
Provides metrics and assessments
Provides feedback
Supervisor
Provides feedback
Displays information
AI4REALNET system
Provides execution plans for implementation
Production dispatching system
Operator
Inputs requests and provides feedback
Prompts information and action options
Provides state at time t
Production information system
Level 1
AI4REALNET System
Provides feedback
Adaptation module
Provides feedback
Assessment module
Provides metrics and assessments
Provides feedback
Regulatory Agent
Supervisor
Provides feedback
Updates Interaction
Displays information
Provides recommendations and scenario information
Human-machine-interaction module
Provides measures to be taken
Execution module
Provides execution plans for implementation
Production dispatching system
Prompts information and action options
Operator
Inputs requests
Provides recommendations
Requests recommendations
Recommen-dation module
Provides scenario information (including events, impact, actions)
Requests prediction for specified scenario
Prediction module
Provides state at time t
Production information system
Level 2
Prediction module
AI agent(s)
Provides state
Provides action(s)
AI4REALNET digital environment
Provides action(s), uncertainty quantifications and explanations
Provides state
Recommen-dation module
Provides scenario outcome
Requests prediction for specified scenario
Evaluation sub-module
Provides run information and measurements
Requests simulation for state and scenario
Simulation engine
Provides state at time t
Production information system

Fig. 8 Hierarchical representation of the systems' building blocks and context

3 Context and Decision Environment in Critical Infrastructures

This section starts with a description of various use cases related to the three considered critical infrastructures, namely, the power grids, the railway network, and air traffic management. Then comes the analysis of context and decision environment related to these critical infrastructures.

3.1 Use Cases Selected by Industry

This section presents the industry-relevant and domain-specific use cases that (a) are focused on critical challenges and tasks of network operators, considering strategic long-term goals, and (b) reproduce real operating scenarios with human operators.

3.1.1 Power Grid

Business Problem Power grids are evolving due to decarbonization and digitalization, integrating clean energy, electrified demand, and new technologies. These changes increase complexity, requiring enhanced control room supervision. Aging infrastructure, automation, and limited developments further challenge operators, who face cognitive overload due to fragmented multiscreen applications. AI can support decision-making by managing complexity, uncertainty, and workload.

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 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. Despite the range of options, their process relies heavily on experience and manual simulation to determine appropriate remedial measures.

Use Case 1: AI-Assisted Congestion Management

Objectives The AI assistant aids TSOs in managing electricity transmission while ensuring (a) safe grid operation and overload mitigation; (b) maximized renewable energy use, reducing thermal dispatch; (c) lower workload for human operators; and (d) enhanced explainability and trust in AI recommendations.

Description The AI assistant monitors the transmission grid using SCADA data and EMS tools, identifying issues and categorizing them for operator intervention. It anticipates congestion, sending alerts with confidence levels while avoiding unnecessary notifications. Action recommendations include topological changes, redispatching, and renewable energy curtailment. Operators review suggestions, explore alternatives, and make final decisions. The AI assistant verifies the decision via load flow calculations and logs interactions for continuous learning.

Human-AI Interaction Human-AI interaction follows a hybrid approach where the operator retains decision-making authority. The AI assistant continuously learns from operator feedback and refines recommendations.

Use Case 2: Transferring AI from Simulation to Real Grid Operations

Objectives Evaluates AI deployment in real-world operations, where actual conditions may differ from training environments. The objectives include (a) ensuring AI effectiveness beyond simulated conditions; (b) strengthening human trust in AI-assisted operations; and (c) enabling iterative refinements through operator feedback.

Description Two operational pathways are explored:

1. Adaptive AI monitoring: The AI assistant continuously assesses the grid, raises alerts, and suggests actions while accounting for uncertainties from noisy or missing data. Operators validate and refine AI recommendations.
2. Human-enhanced AI: When data is incomplete, the AI requests additional input from the operator. Human expertise fills knowledge gaps, refining AI learning and improving recommendations.

In cases where data limitations restrict AI autonomy, it alerts the operator, prompting manual input. The operator's decisions, informed by simulation tools, guide AI forecasting and recommendation assessment.

3.1.2 Railway Network

Business Problem Growing environmental concerns and evolving mobility policies are driving increased demand for railway network capacity, leading to denser traffic and a greater need for efficiency and resilience in railway traffic management. Addressing these challenges requires either significant infrastructure investments or advanced dispatching technologies. AI-based support systems can enhance dispatchers' capabilities by automating certain decision-making processes and assisting human operators in managing complex 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.

Use Case 1: Automated AI-Based Re-scheduling

Objectives The system aims to fully automate railway re-scheduling, ensuring all services are fulfilled while minimizing passenger delays.

Description AI-driven systems dynamically adapt schedules in response to disruptions, such as infrastructure failures or delays. The system evaluates multiple interventions, including modifying train speeds, reordering train sequences, rerouting trains, or adjusting station platforms. This AI assistant continuously optimizes railway schedules in real time, ensuring efficient network use while reducing delays. Human operators oversee the system, adjusting configurations and identifying the need for updates.

Human-AI Interaction The AI system autonomously performs re-scheduling, monitoring real-time train and track states. It identifies required interventions, implements changes, and evaluates outcomes. Since fully automated railway rescheduling is a novel concept, the AI system is introduced as a supervised tool. Human experts evaluate its performance, ensuring reliability and effectiveness. Operators monitor key indicators, including:

- Traffic conditions (e.g., train counts, network bottlenecks)
- Key performance indicators (e.g., current delay statistics)
- AI confidence levels in its recommendations
- The extent of AI-driven interventions (e.g., platform changes)

Based on these insights, human supervisors decide whether to override AI recommendations, reconfigure system parameters, or adjust the AI model.

Use Case 2: AI-Assisted Human Re-scheduling

Objectives This approach leverages AI-based tools to assist human dispatchers in re-scheduling, ensuring all services are maintained with minimal delays.

Description The AI assistant continuously analyzes real-time train and track data, identifying optimal re-scheduling options in response to disruptions. These recommendations are presented to dispatchers in near real time, allowing them to make informed decisions. The system also anticipates future disruptions, assessing their impact on network stability. The AI assistant enhances human decision-making by providing predictive insights, enabling faster and more efficient responses to deviations from planned schedules.

Human-AI Interaction The AI assistant generates re-scheduling options based on real-time data. Dispatchers can:

- Select an AI-recommended action
- Request alternative solutions
- Override AI suggestions and implement their own decisions

Additionally, the AI system supports human learning by providing contextual evidence for or against specific re-scheduling hypotheses. Supervisors continuously assess the AI's performance, refining decision criteria and system parameters as needed.

3.1.3 Air Traffic Management

Business Problem Air traffic density in European airspaces is rising, driven by economic and environmental concerns that push for time- and trajectory-based operations. Increased traffic and operational complexities may lead to excessive tactical air traffic controller (ATCO) workload, threatening ATM system safety and efficiency. Military airspace activations in regions like the Lisbon FIR further restrict upper airspace usage, requiring frequent traffic deviations.

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.

Use Case 1: AI-Assisted Airspace Sectorization

Objectives Automate sectorization decisions to assist or replace ATC supervisors in balancing ATCO workload while ensuring safe and efficient traffic flows.

Description ATC supervisors manually determine how to split and merge airspace sectors based on real-time demands. Sectorization involves horizontal (2D) and/or vertical (altitude) adjustments, often using predefined configurations. However, unexpected events—such as adverse weather, flight emergencies, or ATCO shortages—require non-standard sectorizations. An AI assistant provides recommendations or automates decisions, optimizing horizontal and vertical sector configurations to balance workload while ensuring safety.

Human-AI Interaction The AI system continuously monitors real-time air traffic data, predicts optimal sectorization adjustments, and presents recommendations. Human operators can:

- Evaluate AI-generated recommendations
- Request additional information or explanations
- Accept, reject, or manually adjust AI advisories

Automation levels range from advisory support (where AI suggests actions) to full automation, where AI executes decisions unless overridden by human operators. Logged decisions and interactions enable continuous AI learning and adaptation.

Use Case 2: AI-Assisted Flow and Airspace Management

Objectives Optimize sector capacity and environmental efficiency when military airspace is activated by recommending deviations with minimal impact on sector workload and flight efficiency.

Description When military airspace is activated, ATCOs must reroute flights to avoid restricted areas, impacting sector capacity and fuel efficiency. An AI assistant supports ATC and Flow Management Position (FMP) supervisors by analyzing airspace configurations and suggesting optimal traffic flow adjustments. The system dynamically optimizes routing decisions to minimize en route inefficiencies while maintaining adherence to sector workload limits.

Human-AI Interaction The AI system observes real-time airspace data, forecasts traffic flow disruptions, and recommends airspace and routing adjustments. Human operators can:

- Review AI-generated routing and sectorization recommendations
- Provide feedback and nudge AI-generated decisions
- Modify AI-based configurations based on operational constraints

At lower automation levels, AI assists supervisors by providing recommendations, while higher automation levels allow AI to autonomously execute sectorization adjustments unless overridden. The AI system continuously learns from human feedback to refine its recommendations.

3.2 Decision Environment

Decisions in critical network infrastructure operations are at the heart of operational processes. They can be described in three main points (see Fig. 9). Firstly, they are made to manage constraints on a network capacity that can stem from external events (operational disruptions or emergencies). These events are detected through

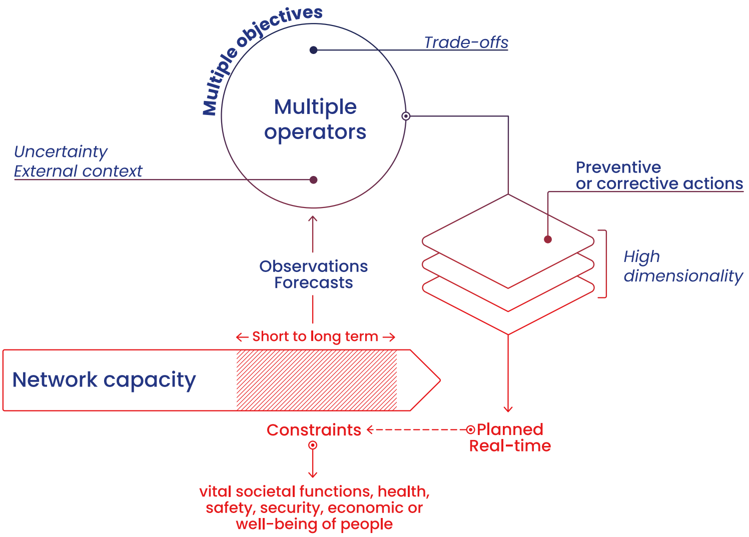


Fig. 9 Decisions in critical network infrastructure operations

observations or forecasts of the state of the infrastructure that include a certain level of uncertainty and an external context. Secondly, they 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.

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 and economic, environmental, and public effects.*

Three industrial domains are considered in this study to extract generic aspects of decisions across critical infrastructures: power grid management, air traffic management, and railway network management. 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. This questionnaire is structured into four main topics: context, characteristics, impacts, and evaluation of decisions. Based on all data collected, a similarity score was computed across pairs of domains to give an idea of the commonalities between the three domains. The results are shown in Table 3.

Table 3 Decision analysis comparison across domains

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%

Even if the decision context is different for each domain (which can be explained by the fact that each domain remains intrinsically different), it could be observed 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 for the development of a holistic conceptual framework.

Then, the second highest similarity is obtained on the evaluation part, which defines a set of KPIs for each domain. The following KPIs were similar across all domains: assistance, relevance, and trust in the AI system. Within the multidomain work, this shows the interest in the evaluation that will be carried out in the future. 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 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, as could be observed, the two most similar pairs of domains are "Railway-Air Traffic" (highest) and "Electricity-Railway." The following section introduces the decision-making process, which is mainly based on observations and analyses of the context.

4 Decision-Making Process in Critical Infrastructures

This section presents the core concept of the conceptual framework, which is the decision-making process in critical systems. It involves a dynamic interplay between human expertise, AI-based decision-making capabilities, and their collaborative interaction. Based on the decision context introduced in the previous section, three key dimensions of decision-making process are explored in the following: human decision-making, which leverages domain knowledge and intuition; AI-based decision-making, which provides data-driven insights and scalability; and human-AI interaction, which integrates these strengths to optimize decisions under complex and uncertain conditions.

4.1 Human Agent and Decision-Making

This section presents the conceptual framework from a socio-technical systems perspective, emphasizing that work systems consist of interacting social and technical subsystems, requiring joint optimization for overall performance. Optimizing one subsystem in isolation may reduce effectiveness due to its interdependence. This leads to two key conclusions: AI design must consider human factors, and the social subsystem—including skills, task design, and organizational culture—must be adapted to effectively utilize AI. For instance, if AI provides recommendations, human decision-makers need both the skills and the autonomy to evaluate them without fear of blame, which could otherwise lead to blind acceptance. While socio-technical design principles are well established [33], further research is needed on AI integration [34–36].

Against this background, three aspects of AI integration in socio-technical systems are examined in the following sections, namely, different design approaches and their effects on human behavior, normative aspects of AI design, and descriptive aspects of AI design. Most of the use cases described in Sect. 3.1 involve interaction between a human operator and an AI agent, making the analysis of human decision-making behavior essential to effectively addressing their objectives.

4.1.1 Different Design Approaches and Their Effects on Human Behavior

The integration of AI into socio-technical systems, particularly in critical infrastructures such as air traffic management (ATM), requires careful design to balance automation and human involvement. Two primary approaches involve *automate*, which replaces human effort, and *informate*, which enhances human capabilities. Full automation is impractical for complex tasks, often leading to unintended negative consequences such as operator fatigue, loss of skills, and overreliance on technology. To maintain system resilience, AI must support human flexibility and decision-making by enhancing their cognitive process rather than relegating humans to passive oversight roles. Research emphasizes the need for AI-human collaboration, where AI design fosters human engagement, autonomy, and motivation. Ensuring that AI systems empower rather than undermine human involvement is crucial for effective performance and safety.

4.1.2 Normative Aspects

Miller [37] outlines and compares five types of explainable AI (XAI) frameworks that support decision-making in human-AI collaboration, ranging from simple AI recommendations to more interactive models where AI explains, interprets, or evaluates human decisions (Evaluative AI). These types influence *human decision-making*, *motivation*, *learning*, and *trust* in AI. In the following, their impact is examined through the scenarios considered for interaction between humans and AI: AI assisting humans, joint human-AI decision-making, and autonomous AI with human supervision.

Human Decision-Making

AI-based decision-support systems primarily rely on recommendations, but research indicates that recommendations alone—even when supplemented with explanations—are often insufficient for effective decision-making [37–39]. Instead, joint human-AI decision-making, which leverages the complementary strengths of both, is necessary [34]. Effective decision-making requires an understanding of human cognitive processes, including biases like the anchoring effect and confirmation bias [38, 40]. Human decision-making is not merely a selection of predefined options but a complex cognitive process that involves sense-making, anticipation, and attention management [41, 42]. To support these cognitive functions, AI should facilitate knowledge of operational processes, process monitoring, situation awareness, and bias mitigation. However, current AI systems often fail to sufficiently support these elements, impacting human performance and decision quality [41, 43].

A simplified representation of human decision-making is illustrated in Figure 10, where the human cognitive processes are summarized as monitoring (observation of the current context or situation); building situation awareness as suggested by Endsley [44], which is to be able to understand the current situation and project to a

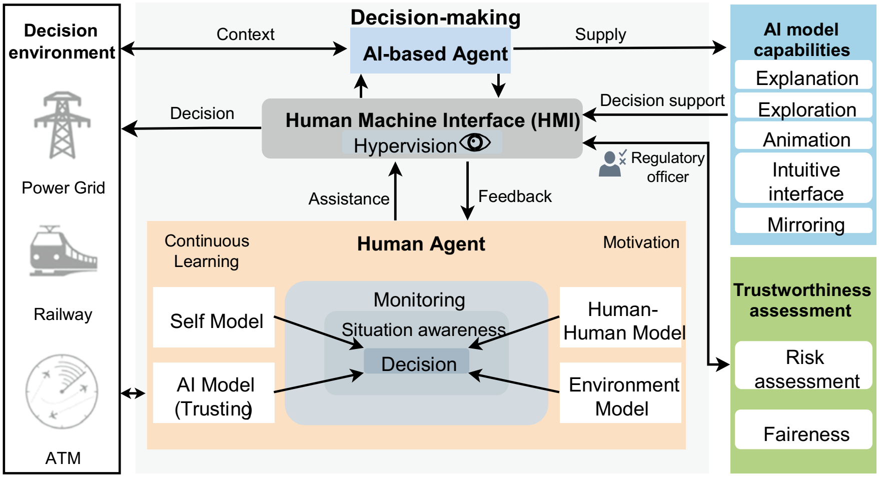


Fig. 10 Decision-making from a human perspective in the proposed conceptual framework involving monitoring, situation awareness, and the decision itself. Human cognitive processes required for decision-making may be supported by AI through explanation, exploration, animation, and mirroring

future state; and finally the actual decision. Beyond the transparency required in AI recommendations (interpretability and explainability), AI can support human cognitive decision-making processes by allowing humans to explore and test their assumptions as well as the implications of their decisions. AI can also support monitoring by requiring humans to animate in detected unusual patterns and mirror the individualized patterns in human behavior to make humans aware of their own biases and variabilities in decision-making. These observations show that AI should support humans in various cognitive processes rather than just making simple recommendations. These are mainly required for the AI-assisted human decision-making use cases, such as AI-assisted congestion management in power grids, AI-assisted human rescheduling for railway networks, and AI-assisted airspace sectorization for ATM.

Human Motivation

Intrinsic motivation is crucial for the effective use of AI, as algorithm aversion and reluctance to adopt IT tools are common [45–47]. AI has an impact on humans' tasks and therefore on their motivation [43]. Task design significantly impacts motivation, which should support the following three critical mental states necessary for intrinsic motivation: meaningfulness, autonomy, and feedback [48]. AI should be designed to enhance these aspects by providing clear explanations, enabling genuine choices, and offering constructive feedback to human, inciting both good work and their usage. In decision-making scenarios, AI recommendations without explanations reduce user engagement, whereas interpretable models and explained recommendations improve transparency and task-related motivation. More advanced AI approaches, such as cognitive forcing and evaluative AI [37], further enhance motivation by increasing user involvement and responsibility in decision-making.

The proposed conceptual framework enables the exploration of different AI-human collaboration models, each affecting intrinsic motivation differently. In an AI-assistant scenario, where AI provides recommendations with or without explanations, user motivation is often low due to a lack of deep understanding, limited autonomy, and missing feedback. Joint human-AI decision-making, including cognitive forcing and evaluative AI, fosters intrinsic motivation by involving users in the decision process, enhancing their sense of responsibility and effectiveness. Conversely, fully autonomous AI, where humans act only as supervisors, poses challenges for motivation and effectiveness due to limited involvement, aligning with [49] irony of automation. In general, AI must be designed to support meaningfulness, autonomy, and feedback to enhance human motivation and performance in AI-assisted tasks.

Human Learning

In addition to the information that could be observed by monitoring and enhanced using situation awareness for decision-making, the human knowledge that may be gained through experience is more enduring. This knowledge could be used as a basis for monitoring and decision-making. Polanyi [50] proposes the distinction between explicit (transferable to others) and tacit (implicit and not transferable) knowledge. This distinction is important because tacit knowledge is the main source of decisions made by domain experts. It may also be interdependent to other individuals' knowledge, forming a community [51]. In this sense, it is important that AI supports humans to earn tacit knowledge through experience.

Four different domains of human knowledge are depicted in Fig. 10 as human mental models, which are the environment model (knowledge about decision context), human-human model (interrelation between the different individuals' work), AI model (knowledge about their capabilities and limitations which could induce the human trust), and self-model (human knowledge about themselves). Human learning is a multifaceted process encompassing psychological, physical, and social dimensions, shaping how individuals perceive and interact with the world [52]. Based on Kolb's [53] Experiential Learning Theory, describing learning as a cyclical process involving four stages (i.e., concrete experience, reflective observation, abstract conceptualization, and active experimentation), AI should also provide the opportunity for the human to learn through exploration, animation, and mirroring. This could be helpful for all the use cases of critical infrastructures (see the previous section), where the human intervention as a supervisor is enhanced using the AI assistance.

Human Trust

Human trust in AI is a dynamic relationship influenced by beliefs, knowledge, emotions, and experiences with the AI and with their mental AI model as shown in Fig. 10 [54, 55]. Unlike trustworthiness, which is a static attribute of a particular AI, trust evolves with repeated interactions [56]. As outlined by Parasuraman and Riley [57], appropriate trust requires users to understand AI's capabilities and limitations to avoid both overreliance (where human blindly approves the AI recommendation) and under-utilization (not having the opportunity to experience and develop trust). On the other hand, the transparency of AI models should not only provide explanations but also allow users to explore and test AI functionality through direct experience and develop the appropriate trust accordingly [58].

The trust is examined across three previously mentioned human-AI collaboration scenarios considered in the conceptual framework. The AI-assistant to human model provides recommendations but limits user involvement, often leading to overreliance or rejection of AI decisions [41, 49]. Joint human-AI decision-making, which integrates cognitive forcing and Evaluative AI [37], fosters trust by increasing transparency, interaction, and feedback opportunities. Autonomous AI, where humans act as supervisors, presents the greatest challenge for trust development due to minimal human engagement. Ultimately, fostering appropriate trust requires AI systems to prioritize transparency, interactive exploration, and feedback mechanisms, ensuring that trust is informed by hands-on experience and a clear understanding of AI's strengths and limitations.

Acceptance

Initial acceptance of new technology is crucial for its successful deployment, as rejection can occur even before use, creating a paradox where trust develops through experience but may not form without initial engagement. Research from sociology, psychology, and information systems highlights human-technology compatibility as a key factor in overcoming this hurdle, with acceptance increasing when a system aligns with users' values, experiences, and needs. The compatibility exists at various cognitive levels, from basic handling to decision-making strategies. Studies in air traffic control (ATC) show that strategic conformance—where AI solutions resemble human decision-making—enhances acceptance, as seen in higher approval of personalized recommendation systems over generalized ones. However, while strategic conformity aids initial adoption, its long-term benefits may diminish with prolonged human-AI interaction [59].

4.2 AI-Based 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 responsibly and effectively.

Figure 11 presents the logical view of human-AI decision-making, emphasizing the role of the AI-based agent. Both AI and human agents observe the context and decision environment, but the AI-based agent excels in processing and analyzing complex contextual information in real-time—capabilities beyond human capacity. Leveraging various strategies and methods, such as knowledge-assisted AI, meta-awareness for AI assistants, human-AI co-learning, and multi-objective learning, the AI-based agent anticipates events and provides recommendations to the human agent. For effective integration into the proposed conceptual framework, AI-based models must exhibit specific characteristics that facilitate seamless interactions

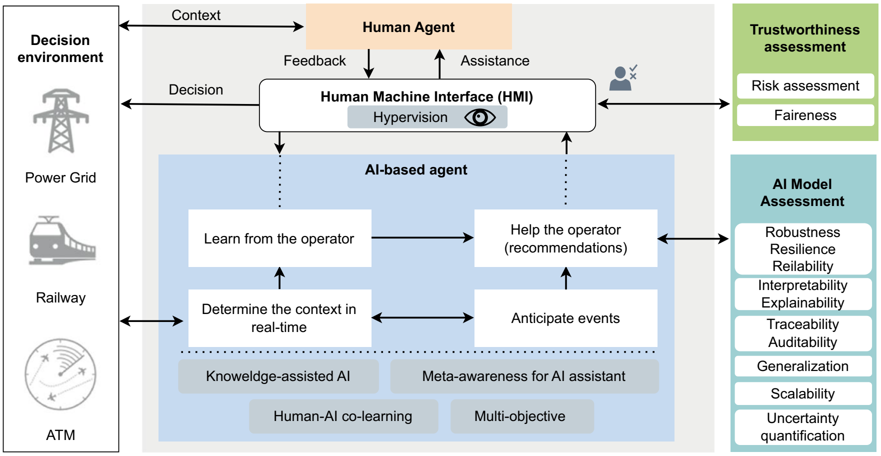


Fig. 11 Decision-making from an AI perspective in the proposed conceptual framework

between AI and human decision-makers across different scenarios and interaction modes. Additionally, the AI-based agent should continuously learn from human feedback and incorporate it into future recommendations. Ultimately, any decision must align with predefined trustworthiness KPIs to ensure reliable implementation in the decision environment. The following sections provide a detailed discussion of the AI-based agent's characteristics, along with the methodological and algorithmic aspects of AI-based models.

4.2.1 Characteristics of AI-Based Decision-Making

In the context of human-AI decision-making process of a proposed conceptual framework, AI systems can augment human judgment by providing data-driven recommendations or by increasing their cognitive process. This helps enhance efficiency and reduce bias in critical decisions. However, this synergy also brings challenges for which AI-based assistance should adhere to a set of properties (see Fig. 11).

They should be robust and resilient to perturbations and maintain reliability with respect to what they have learned during the training phase. Their assistance must be explainable to human operators, with an architecture that is easily interpretable. Additionally, generalization and scalability are essential for deploying these solutions in real-world conditions. Lastly, their assistance should be accompanied by clear indications of uncertainty levels, increasing the trust of the human operator. These properties are discussed in greater detail in the following sections.

Robustness, Reliability, and Resilience

Robustness in AI can be examined from two perspectives: technical robustness and social robustness. *Technical robustness* refers to a system's ability to sustain its performance despite natural or adversarial perturbations (ISO/IEC 24029-2). It can be assessed locally (for specific inputs) or globally (across all inputs). Evaluation methods include measuring sensitivity—how output changes with input variations—or introducing adversarial perturbations to test system stability [60]. Metrics like output variance or reward function deterioration can quantify robustness. Furthermore, different learning strategies may impact the robustness, e.g., online learning impacts robustness, requiring continuous test-time monitoring. *Social robustness*, on the other hand, ensures that the AI system duly considers the context and environment in which it operates, guided by frameworks like ALTAI. Digital environments associated with each critical infrastructure may help assess social impacts, such as carbon emissions reduction.

Reliability in AI refers to its ability to perform as expected, even with novel inputs (EU-U.S. Terminology and Taxonomy for AI). Unlike robustness, which considers AI performance under disruptions, reliability focuses on stable performance within a consistent data distribution [61]. It relates closely to generalization, ensuring AI models perform similarly across different test datasets. Estimating epistemic uncertainty can also help assess reliability by correlating performance with uncertainty levels. Reliable AI systems should demonstrate consistent results across standard operational conditions, independent of external disturbances.

Resilience is the AI system's ability to prepare for, withstand, and recover from unexpected perturbations or attacks (EU-US Terminology and Taxonomy for AI). 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 12 depicts a conceptual definition of resilience quantification for a reward function. In this scheme, resilience can be quantified by (a) the gray area between the reward curves of the unperturbed and perturbed AI system, (b) the minimum reward in the degradation state and the maximum reward in the restorative state, and (c) the duration of the degradation and restorative stages. These metrics should be

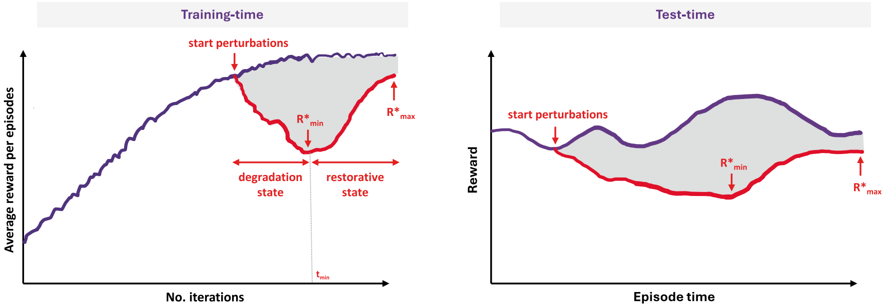


Fig. 12 Concept of resilience quantification in training-time and test-time phase

computed for different probability levels of the perturbation or by defining a maximum number of perturbations or a perturbation budget.

Interpretability and Explainability

Explainability measures the capability of a human user to understand how models make predictions or decisions, where the model's transparency (interpretability) is a way to support explainability [37]. According to Molnar [62], explainability is contextualized by a specific input, and it often requires additional information, which the decision model does not generally generate. XAI techniques refer to the set of methods that aim to generate local explanations for black-box models, e.g., artificial neural networks predictions.

To ensure a thorough evaluation, Vouros [63] proposes a set of additional human-related metrics evaluating the interaction between a human and the explanation. These metrics 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.

Traceability and Auditability

Traceability refers to the process of establishing a clear and direct connection between the stakeholders' requirements and the product developed in the context of software development. When applied to AI systems, this connection covers the design elements, code implementation, test 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.

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 [64]. This concept is pivotal in building human trust in AI systems. One might argue that, when deploying AI into real-world systems, auditability is as important as model performance.

In the proposed conceptual framework, auditability and traceability serve three main purposes. First, they ensure human control and safe operation by detecting changes in RL elements, maintaining task consistency, and preventing external interference. Second, they support regulatory compliance, monitoring, and inspection of AI and human inputs. Third, they enable tracking of both AI recommendations and human influence on decisions. Additionally, they help identify performance issues, improving system maintainability.

Generalization

Generalization, in the context of AI, is the ability of a trained model to perform well on previously unseen data that is derived from the same distribution as the data explored during the training phase. For RL, Nichol et al. [65] consider an agent to generalize well when it can adapt to previously unencountered situations drawn from the same task explored during the training phase, such as varying difficulty levels in a simulation or different configurations in an operational environment. As in RL, the policy is optimized based on a reward signal, which is often formulated ad hoc for specific scenarios. The generalization capability may be limited by the overfitting problem [66], where an agent is trained to maximize the cumulative reward in a fixed scenario. The trade-off between generalization and overfitting is still an open challenge [67]. The problem of generalization in RL should be addressed on three different levels: domain diversity (exploration of heterogeneous environment configurations), exploration-exploitation trade-off (find a balance to optimize on various experiences to prevent overfitting) [68], and finally reward engineering (ensuring that reward formulation is not overly tailored to a specific scenario).

In the context of the proposed conceptual framework, it is required to ensure that the AI model is robust and capable of generalizing 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. The generalization capability of a model could be measured by observing the change in the reward/loss of the model when visiting novel data.

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 [69] and Ulanov et al. [70], where they assess the scalability of an AI distributed model by measuring the empirical speedup obtained from a system while increasing the computational resources.

The conceptual framework describes an RL-powered AI agent supporting human operators working in control rooms to maintain a critical infrastructure operational. When a disruption occurs, operators usually have a tight time frame to analyze the situation and devise a remedial plan to return to safe conditions. Consequently, the computational overhead of the AI agent, as well as the additional workload resulting from its suggestions to the operator, must be kept minimal regardless of the scale and complexity of the problem.

As a result, the scalability challenge is twofold. On one hand, from an engineering perspective, an AI decision-making model should scale based on hardware availability. On the other hand, from a theoretical RL perspective, the system's effectiveness and performance should not be compromised by the complexity resulting from the combinatorial nature of Multi-Agent RL (MARL) with an arbitrary number of agents [71]. The combinatorial complexity can be mitigated by factorizing the learning process across multiple agents, enabling each agent to observe and make decisions simultaneously within a shared environment.

Uncertainty Quantification

Uncertainty quantification (UQ) systematically characterizes and manages uncertainties inherent in both AI models and real-world data, and it is crucial when an AI agent supports the decision-making of a human operator for critical infrastructures [72, 73]. In human-AI interactions, UQ ensures decisions are reliable and robust by addressing different aspects of uncertainty. Epistemic uncertainty arises from insufficient training data, leading to gaps in the model's knowledge and limiting its ability to generalize to unseen scenarios. Quantifying epistemic uncertainty helps humans interacting with an AI system better understand AI limitations, which is crucial for the conceptual framework integrating a human-in-the-loop strategy. Aleatoric uncertainty is intrinsic to the data, as it stems from noise, incompleteness, or inaccuracies in input data. Since the proposed conceptual framework is designed for critical infrastructures, environments should support aleatoric uncertainty estimation, particularly from external sources unobservable by an AI agent, such as weather conditions, to ensure more reliable AI-assisted decision-making.

Beyond these technical dimensions, UQ plays a crucial role in decision-making under uncertainty, offering probabilistic frameworks to evaluate risks and outcomes for critical infrastructures. This helps human operators make informed decisions despite uncertainty. Human-AI collaboration is also enhanced, as transparent uncertainty metrics allow humans to interpret AI predictions with greater awareness of their limitations. For instance, AI-generated decisions are augmented with confidence levels in power grid use cases, ensuring that human operators understand the system's constraints. Finally, UQ contributes to resilience and reliability by ensuring AI systems are not only accurate but also aware of their limitations. This awareness leads to more robust designs and operational strategies, improving the overall dependability of critical infrastructures.

4.2.2 AI-Based Methods and Strategies

This section explores four key methodologies and strategies employed in the design of AI-based models. These models play a crucial role within the proposed conceptual framework, ensuring the fulfillment of objectives related to the analyzed critical infrastructures.

Knowledge-Assisted AI

Knowledge-assisted AI integrates preexisting knowledge with data-driven methods, often resulting in hybrid or neuro-symbolic approaches [74, 75]. This knowledge can come from heuristics, symbolic rules, or physics equations, enhancing generalization, particularly in data-scarce environments. While such knowledge improves transparency and interpretability, it may not cover all possible scenarios, necessitating data-driven elements to fill gaps, resulting in knowledge-assisted AI approaches. In the closely related area of "informed machine learning," key classification factors of existing approaches include the source, representation, and integration of knowledge [75].

The primary goal of knowledge-assisted AI is to enhance system performance rather than add new functionality, making AI systems more robust, interpretable, and generalizable. These are of particular importance in the considered critical infrastructures, and evaluation could specifically target these aspects by, for example, testing the system under abnormal conditions. Knowledge structured in an accessible format helps establish cross-domain applicability, whereas more exotic formats could be useful in domain-specific applications. The type of knowledge that could be considered varies from one domain to another. For example, in power grids, the physics constraints or expert knowledge could be exploited by the models to provide a more robust and compliant set of actions.

While various approaches to knowledge-assisted AI systems have been explored, fewer studies have directly examined their impact on decision-making. Decision-making in AI is primarily studied in reinforcement learning (RL), which can be categorized into model-based and model-free methods. Model-based approaches learn an explicit model of the environment, typically using supervised learning techniques, and could, therefore, benefit directly from neuro-symbolic and informed machine learning methods surveyed by van Harmelen and Teije [74] and Von Rueden et al. [75].

In contrast, model-free methods do not rely on an explicit model but can still incorporate different types of knowledge. For example, prior works have leveraged logical rules for high-level transitions [76–79], spatial invariances [80], knowledge graphs [81], differential equations [82], and human feedback [83–85].

Meta-Awareness for AI Assistants

Meta-awareness in AI assistants is essential for effective human-AI teaming, where AI systems complement human operators by handling large data volumes while humans manage unforeseen and edge-case situations [41]. Unlike ML-based systems that lack causal reasoning, AI assistants must anticipate events, manage uncertainty, and flag anomalous situations [86, 87]. To enhance reliability, AI systems should quickly learn from failures, alert users to high-uncertainty states, and ensure meaningful human control [34], 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 prompts them to seek human assistance (e.g., send alarms to the operator when the proposed actions are of low confidence).

In the context of the proposed conceptual framework, the meta-awareness framework considers the following phases: (1) monitoring infrastructure with contextual knowledge extraction [88], (2) predicting system behavior and anomalies using uncertainty quantification, and (3) transferring control to humans when uncertainty is high [89]. Deferral mechanisms, such as those in Bondi et al. [90], enable AI to decide when to defer to human judgment based on uncertainty levels and operational context, ensuring reliable AI-assisted decision-making.

Figure 13 illustrates a prototype of a deferral mechanism that, following the nomenclature proposed by Bondi et al. [90], 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

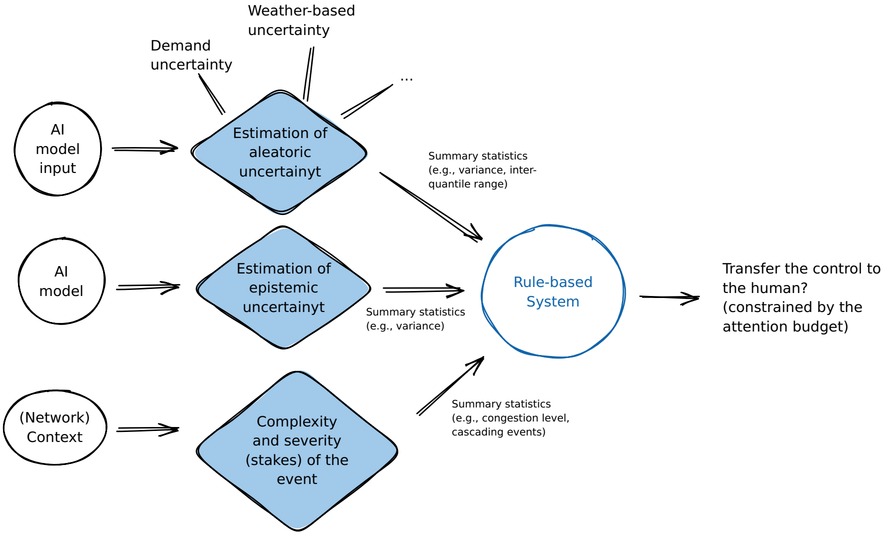

Weather-based uncertainty
Demand uncertainty
...
AI model input
Estimation of aleatoric uncertainyt
Summary statistics (e.g., variance, inter-quantile range)
AI model
Estimation of epistemic uncertainyt
Summary statistics (e.g., variance)
Rule-based System
Transfer the control to the human? (constrained by the attention budget)
(Network) Context
Complexity and severity (stakes) of the event
Summary statistics (e.g., congestion level, cascading events)

Fig. 13 Prototype schematic of a deferral mechanism that learns to defer decision-making from the AI model to a human

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.

Human-AI Co-learning

Human-AI co-learning, often referred to as human-AI teaming or human-machine collaboration, aims to establish a continuous and mutual learning process between humans and AI systems. Unlike traditional AI-human interactions, co-learning focuses on leveraging the strengths of both agents while mitigating their respective weaknesses, leading to superior team performance.

A preliminary design for a co-learning AI agent is inspired by the work of van den Bosch et al. [91], which outlines six essential models that an AI system must develop and refine to enable effective collaboration. These include the taxonomy model (shared language), team model (work agreements and hierarchy), task model (knowledge about tasks and strategies), self-model (AI's internal state), Theory-of-Mind model (understanding human agents' inner states), and communication model (exchange of information based on shared understanding).

These models collectively enable the AI agent to align with human cognition, adapt dynamically, and foster productive interactions within the team.

As illustrated in Fig. 14, these models interact through structured information flows, ensuring that AI agents can process human communication, infer behavioral cues, and communicate their internal states transparently. This approach provides a conceptual foundation for human-AI co-learning, offering insights into necessary functionalities without yet specifying technical implementations.

An example of co-learning is AI-assisted human rescheduling in railway operations, where human and AI capabilities could be improved in parallel. Human feedback could be used to improve future AI recommendations, and AI systems support human learning by providing contextual evidence for or against specific rescheduling hypotheses.

Multi-Objective Reinforcement Learning

Designing multi-objective AI agents requires addressing both training and operational phases while integrating human preferences. In training, the AI optimizes a total reward function, $\boldsymbol{R}_{\mathrm{tot}}$, which aggregates individual objectives through a scalarization function ($\boldsymbol{R}_1$, $\boldsymbol{R}_2$, …, $\boldsymbol{R}_n$) [92]. Without human input, this function must be predefined, though a more flexible but complex alternative would be training an agent to handle multiple reward combinations dynamically.

During operation, AI-generated solutions are ranked by $\boldsymbol{R}_{\mathrm{tot}}$ and presented to human operators, who may override the AI's ranking based on their expertise, implying an implicit preference function different from the predefined function $\boldsymbol{U}$. To refine AI decision-making, a feedback loop is introduced, capturing

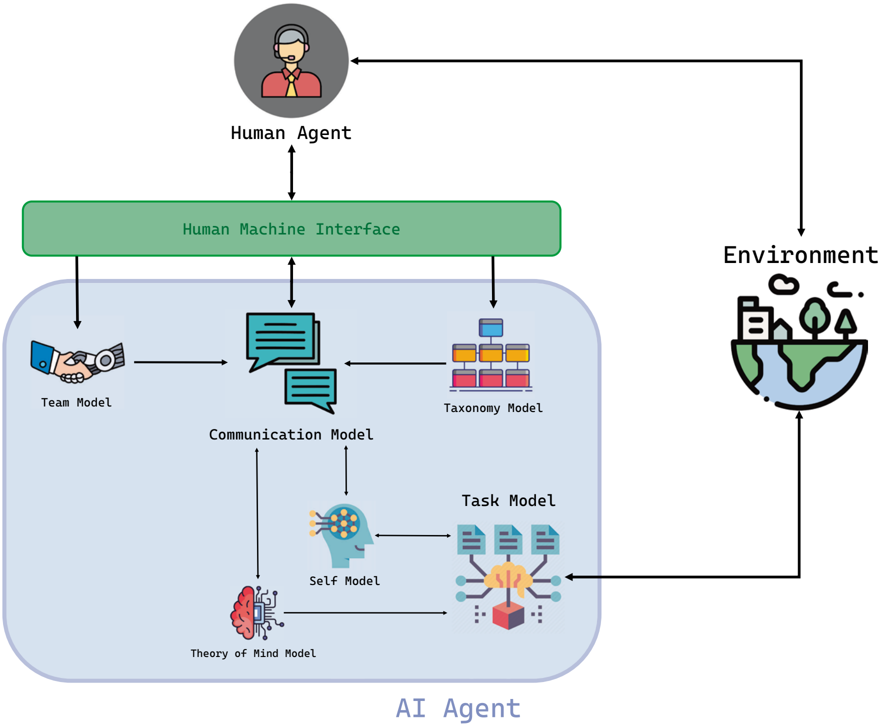


Fig. 14 Descriptive schematic of a co-learning AI agent

discrepancies between AI recommendations and human choices. Ideally, this data would be used to adjust the AI's reward function to better reflect expert preferences, although perfect alignment may be unattainable. Instead, heuristic methods developed in collaboration with experts can help approximate a suitable *U*-function.

A roadmap for developing multi-objective AI agents emphasizes a progressive approach, beginning with heuristic models and visualization tools (e.g., spider charts) to enhance explainability and transparency before integrating human feedback. The final step involves refining the utility function through iterative learning. Figure 15 illustrates a multi-objective visualization in a power grid congestion management use case, where operators select among AI-generated actions based on multiple conflicting objectives. Recording operator choices inform future AI adjustments, fostering alignment between human and AI decision-making.

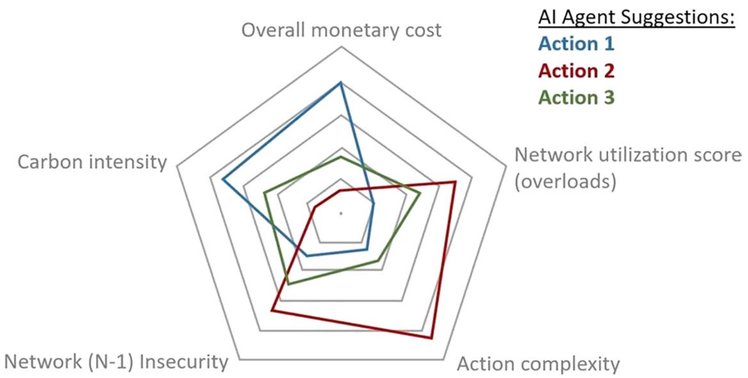


Fig. 15 Example of multi-objective visualization

4.3 Human-AI Interaction in Critical Infrastructures

Within the conceptual framework and in the context of human-AI interaction, 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 Fig. 20). At each stage, the levels of automation consider the division of roles and responsibilities between humans and machines 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 following sections describe in greater detail the interaction scenarios, the design steps based on existing frameworks, and the hypervision tool, which provides the human operator with real-time insights, system diagnostics, and performance analytics, enabling better oversight and informed decision-making.

4.3.1 Interaction Scenarios

This section describes the three interaction levels between humans and AI that can be leveraged within the proposed conceptual framework based on the observed context.

Human in Full Control

This 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

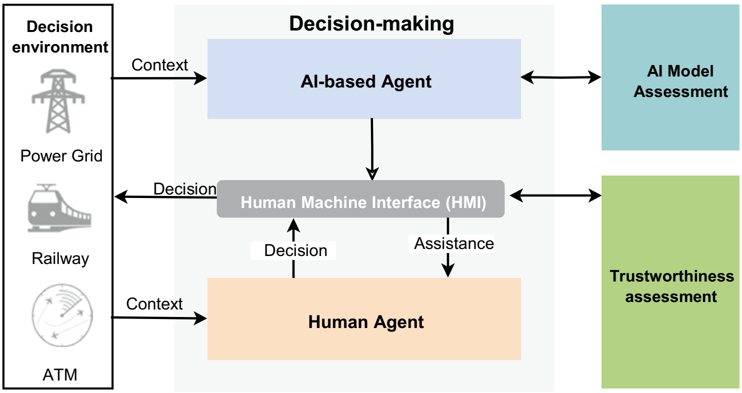


Fig. 16 Human in full control scenario

decision-making power rests with humans. Overall, maintaining human control in AI interactions ensures that technology serves to augment human capabilities (cognitive processes) while safeguarding against unintended consequences or misuse. The logical view corresponding to this mode of interaction is shown in Fig. 16. In addition to the observed context, the digital environments provide us with a set of tools to simulate real scenarios, enabling the assessment of the decision's impact before its application in a real-world context. When an event occurs, human operators should take some actions (decisions) to keep the environment in a stable state. They could interact with AI assistance to enhance their capabilities (exploration) at the decision-making step. The AI assistant may also provide explanations to guide human operators in the 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. This mode of interaction is required for both use cases of the power grid domain and the airspace sectorization assistant in ATM.

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 Fig. 17), 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 socioeconomic and environmental factors, make

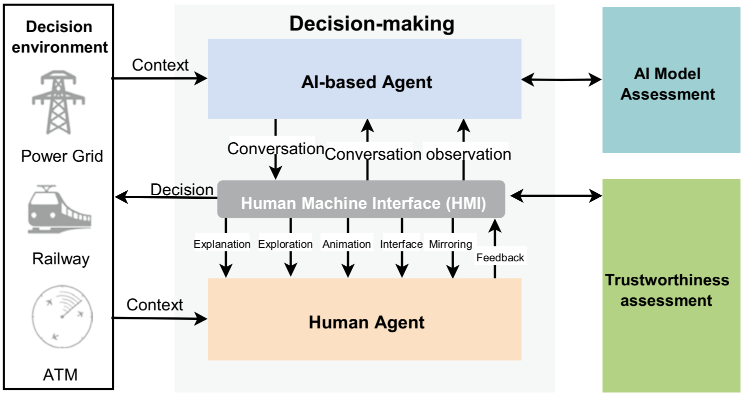


Fig. 17 Human-AI co-learning scenario

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 and cognitive processes. 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. This mode of interaction is required by the AI-assisted human rescheduling in railway operation and flow and airspace management assistant in ATM.

Human as Supervisor

Autonomous AI systems with human supervision (see Fig. 18) in the context of critical infrastructure refer 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

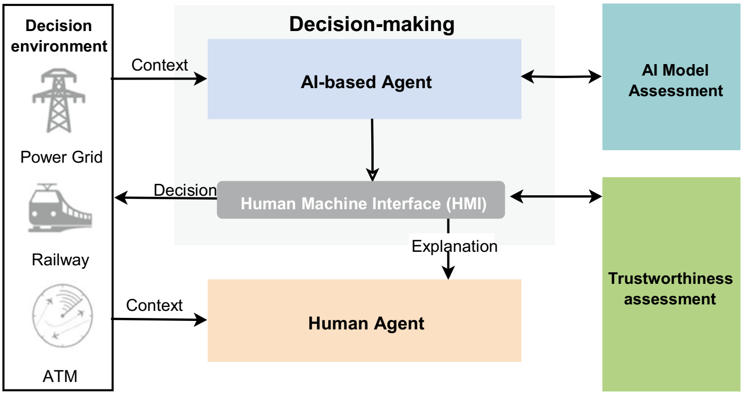


Fig. 18 Human as supervisor or autonomous AI scenario

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. This is required by the automated rescheduling in railways operations, where a human as supervisor ensures the reliability and effectiveness of AI-based decisions and could override AI recommendations.

4.3.2 Describing and Designing Human-AI Interaction

For describing and designing human-AI interactions, cognitive engineering offers insights from human-automation teamwork, emphasizing shared control, autonomy, and transparency. However, a universal design framework is lacking. The proposed conceptual framework explores integrating two complementary frameworks: JCF [25, 94], which focuses on planning and executing activities among agents, and Ecological Interface Design (EID) [95], which enhances transparency by visualizing constraints. Both are rooted in Cognitive Systems Engineering (CSE), which prioritizes designing interactions based on the work environment rather than specific agents. Figure 19a illustrates this triadic approach, where JCF structures activities, while EID ensures clarity in decision-making by depicting system constraints. Combining these frameworks allows for a structured approach to human-AI collaboration, where EID dictates what information to display and JCF determines when and how it should be presented to support decision-making. Figure 19b represents this integration, showing how AI can enhance perception and action leverage at different abstraction levels, ensuring alignment between what is seen, decided, and executed.

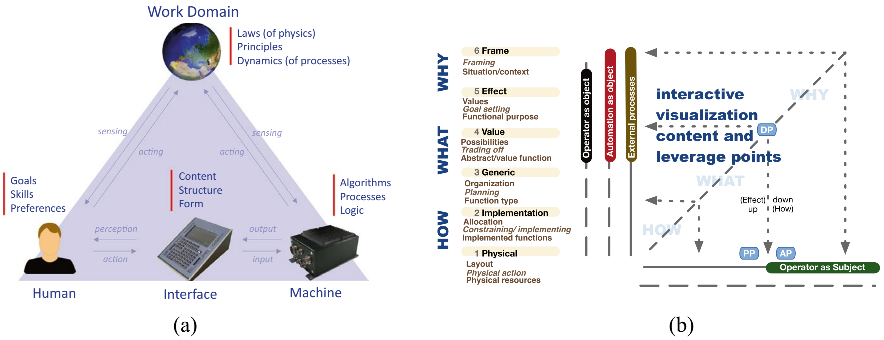


Fig. 19 Designing human-AI interaction: toward a common framework. (**a**) Triadic approach to human-AI interaction. (**b**) Merger of JCF and EID on a functional level

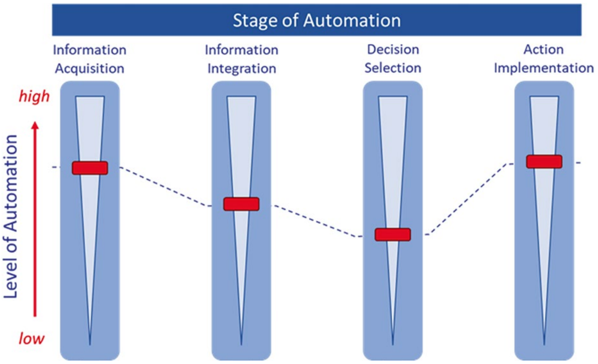


Fig. 20 Stages and levels of automation modeled after human information processing steps [93]

The proposed conceptual framework considers three human-AI teamwork configurations: AI-assisted human control (human in control), joint human-AI decision-making (co-learning), and autonomous AI (human as supervisor). These scenarios align with the established "stages and levels of automation" model (Fig. 20), which determines how autonomy and authority are distributed between humans and AI across four stages: information acquisition, integration, decision selection, and action implementation. In AI-assisted control, AI supports human decision-making but does not execute actions. In joint decision-making, AI and humans work in parallel, learning from each other's actions. In autonomous AI, the system operates independently, with human intervention reserved for system failures. The appropriate level of automation depends on operational context and system capabilities, meaning a universal model does not exist and must be tailored to specific applications. Figure 20 links these automation levels to the interface design from Fig. 19b, showing how AI's role in decision-making evolves based on its autonomy level.

4.3.3 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 [96].

Hypervision aims to deliver the right information to the right person at the right time while tracking user progress for each task [97]. It provides a unified interface that synthesizes key information and centralizes real-time business events to support decision-making and task prioritization. Operators use Hypervision to understand the operational context, diagnose alerts, and implement solutions efficiently.

By enhancing event prioritization and syncretization, hypervision goes beyond real-time monitoring to anticipate future tasks through forecasting. This shift moves the focus from alarm management to proactive task completion (see Fig. 21a).

Hypervision is structured into four layers as shown in Fig. 21b: *Synthesis*, which integrates data from various online tools; *Formatting*, which determines the best

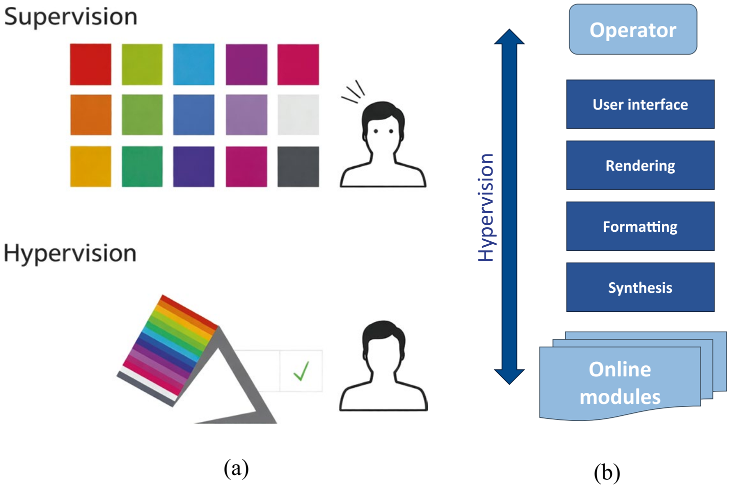
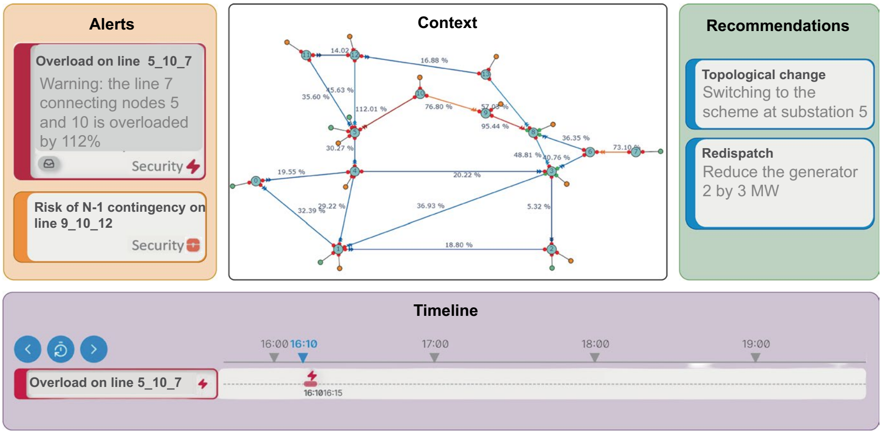


Fig. 21 Hypervision, providing unified interface that synthesizes key information and centralizes real-time business events to support decision-making and task prioritization. (**a**) From supervision to hypervision. (**b**) Hypervision implementation

way to present information (text, tables, graphs, etc.); *Rendering*, which connects formatted data to context (e.g., visualizing line overloads on a map); *User Interface*, which supports synthesis, prioritization, human-machine interaction, and collaborative decision-making.

As an example of hypervision, Amokrane-Ferka et al. [98] propose a virtual bidirectional assistant to support augmented decision-making in complex steering systems (Cockpit and Bidirectional Assistant (CAB) project[1]). This assistant continuously learns from real-time data flows and human decisions, enabling seamless interaction between human experts and AI. As can be seen in Fig. 22, the interface follows a hypervision-based framework with multiple panels: a context panel for real-time environmental visualization, a timeline panel for tracking past events, an alerts panel for notifying operators of risks and system changes, and a recommendations panel that provides AI-driven suggestions. Operators can choose to follow these recommendations based on their expertise and the complexity of the situation.

A major focus of this approach is enhancing the explainability of AI-generated recommendations to support human decision-making. The assistant dynamically assesses the operator's profile and cognitive workload to tailor the flow of information, ensuring optimal handling of complex or atypical situations. The system aims to improve situation awareness, reduce cognitive overload, and enhance overall efficiency by adapting recommendations and interactions to the operator's needs.



Alerts

Overload on line 5_10_7

Warning: the line 7 connecting nodes 5 and 10 is overloaded by 112%

Security

Risk of N-1 contingency on line 9_10_12

Security

Recommendations

Topological change

Switching to the scheme at substation 5

Redispatch

Reduce the generator 2 by 3 MW

Timeline

Overload on line 5_10_7

Fig. 22 Example of hypervision interface (CAB project)

[1] https://www.irt-systemx.fr/en/projets/cab/

5 Epistemological and Philosophical Foundations of Trustworthy AI and Their Adaptation for the Conceptual Framework

This subsection investigates the epistemological and normative foundations of the notion of Trustworthy AI (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 foundation, 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—or at least non-purely quantitative—nature.

5.1 The Epistemological and Normative Grounds of the Notion of TAI

The notion of TAI plays a central role in ensuring AI systems are developed and deployed responsibly, particularly within the European Union's ethics-based regulatory framework. According to the European Commission's Ethics Guidelines for Trustworthy AI, TAI must be lawful, ethical, and robust throughout the system's life cycle. However, there is ongoing debate about what constitutes trustworthiness in AI and whether it is feasible for all systems. Explainability, for example, is widely considered essential for TAI but remains challenging in many AI applications. Furthermore, some critics argue that the concept of TAI is more of a marketing strategy than a meaningful ethical framework, as AI lacks the motivations and moral obligations that are fundamental to interpersonal trust [99, 100]. Consequently, some scholars question whether attributing trustworthiness to AI is conceptually valid or simply a form of ethics washing.

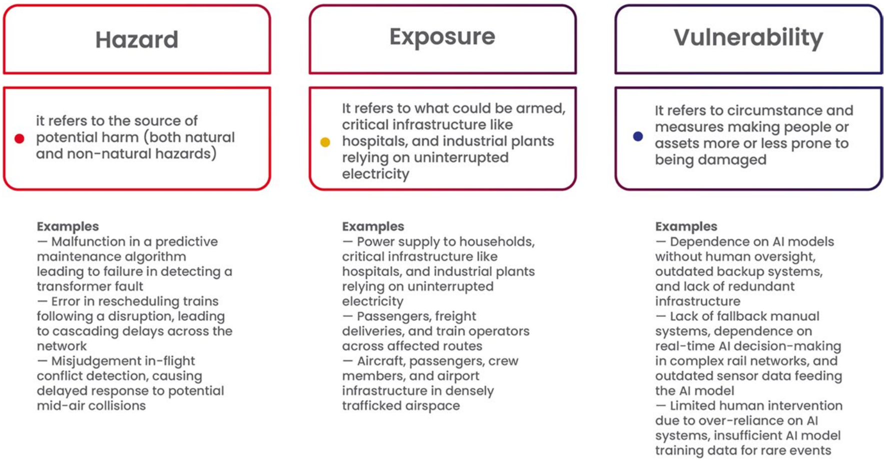
Despite these criticisms, in the design of the conceptual framework, the value of applying the notion of trustworthiness to AI is substantial, as it integrates both technical reliability and ethical acceptability. This aligns with the European Ethics Guidelines for Trustworthy AI and the ALTAI framework, which the conceptual framework uses to assess risks, establish requirements, and validate AI systems. However, to avoid conceptual errors, a distinction is made between trustworthiness in human-human (H-H) and human-AI (H-AI) interactions [101]. While both H-H and H-AI trust involve reliability and ethics, they differ in how ethical considerations manifest. Human trust relies on goodwill and moral obligations, whereas AI trustworthiness is based on compliance with ethical requirements. For example, in a use case where an AI assistant is transferred from simulation to real-world operation, AI must be designed to prevent human manipulation, such as misleading feedback or misuse of the AI learning process. By acknowledging these differences, it is ensured that TAI remains a practical and meaningful framework for developing ethical and technically robust AI systems.

5.2 AI-Related Risk and Uncertainty

The notion of TAI not only integrates crucial aspects of AI system design, deployment, and assessment but also plays a key role in addressing AI-related risk. Traditionally, trust and trustworthiness have been associated with situations of risk and vulnerability [102], making risk assessment a central concern for the conception of the framework. However, risk itself is a complex concept with no universally accepted definition. The classic definition by the Royal Society [103] emphasizes probability, while modern interpretations define risk as the combination of an event's probability and its consequences [104]. This perspective is also reflected in the AI Act, which considers risk as the likelihood and severity of harm (Art. 3, 2). However, the AI Act does not further articulate how risk should be assessed or mitigated, making it necessary to refine the methodological foundations of AI-related risk assessment.

5.2.1 The Components of Risk

A structured approach to risk assessment, often used in disaster risk management, breaks risk into three key components: hazard, exposure, and vulnerability. Hazard refers to the source of potential harm, such as system malfunctions, and is typically assessed through probabilistic estimates. Exposure considers the entities—people or material assets—that could be affected by the hazard. Vulnerability refers to the factors determining how susceptible these entities are to harm. Risk arises from the interaction of these three components, meaning even a low-hazard system can pose significant risks if exposure or vulnerability is high. Conversely, high hazard levels do not automatically translate into high risks if exposure and vulnerability are minimal. This multi-component analysis, commonly used in natural risk management, is also applicable to technological risks, including AI-related risks. As illustrated in Fig. 23, different AI systems may present risks due to varying levels of these components, enabling targeted mitigation strategies such as restricting access to AI-based services or reducing user vulnerability.



Hazard

it refers to the source of potential harm (both natural and non-natural hazards)

Examples
— Malfunction in a predictive maintenance algorithm leading to failure in detecting a transformer fault
— Error in rescheduling trains following a disruption, leading to cascading delays across the network
— Misjudgement in-flight conflict detection, causing delayed response to potential mid-air collisions

Exposure

It refers to what could be armed, critical infrastructure like hospitals, and industrial plants relying on uninterrupted electricity

Examples
— Power supply to households, critical infrastructure like hospitals, and industrial plants relying on uninterrupted electricity
— Passengers, freight deliveries, and train operators across affected routes
— Aircraft, passengers, crew members, and airport infrastructure in densely trafficked airspace

Vulnerability

It refers to circumstance and measures making people or assets more or less prone to being damaged

Examples
— Dependence on AI models without human oversight, outdated backup systems, and lack of redundant infrastructure
— Lack of fallback manual systems, dependence on real-time AI decision-making in complex rail networks, and outdated sensor data feeding the AI model
— Limited human intervention due to over-reliance on AI systems, insufficient AI model training data for rare events

Fig. 23 AI-related risk and its components

5.2.2 Uncertainty

Risk assessment often assumes that potential outcomes can be assigned precise probabilities, yet this is rarely realistic, especially in AI-related scenarios [105]. While probabilistic models, including second-order probabilities and probabilistic intervals, can sometimes quantify uncertainties, large-scale deployments of innovative technologies pose additional challenges. In many cases, historical data is insufficient to inform probability estimates, leading to deep uncertainty where even approximate probabilities are difficult to assign [106]. To address this, AI risk assessment must go beyond the design phase and involve continuous evaluation in real-world contexts. The tool provided in the following pages is designed for ongoing assessments, ensuring that AI systems remain aligned with safety and ethical considerations throughout their life cycle.

5.3 *A Non-calculative Tool for Risk Assessment in Safety-Critical Systems*

In the design of the conceptual framework, the multi-component analysis of risk serves as the foundation for applying the ALTAI framework to a specific context. ALTAI is structured around seven key requirements central to TAI: Human Agency and Oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Diversity and Fairness, Societal and Environmental Well-being, and Accountability. However, in the case of considered safety-critical systems, four of these requirements—Human Agency and Oversight, Technical Robustness and Safety, Societal and Environmental Well-Being, and Accountability—are particularly relevant. These requirements align well with the multi-component risk analysis and help address key challenges, such as the risk of overreliance on AI systems. For example, Requirement #2 (General Safety) from ALTAI, which asks whether threats to an AI system have been identified, can be refined by considering hazard (the probability and impact of threats), exposure (the system's susceptibility to threats), and vulnerability (measures taken to reduce susceptibility). Similarly, the requirement related to Societal and Environmental Well-Being is refined by assessing de-skilling risks in terms of affected skills, workforce, and vulnerability.

In addition to adapting ALTAI's questions, a requirement on risk acceptability is proposed, recognizing that acceptable levels of risk depend on context and available alternatives. Unlike ALTAI, which primarily functions as a tool for ex post self-assessment, the proposed approach reconceives ALTAI's questions as positive requirements to be applied during the design phase. This shift encourages proactive responsibility in AI development, ensuring that safety-critical systems are assessed for risk mitigation from the outset rather than as an afterthought. The key requirements, tailored for the conceptual framework's specific needs, are summarized in Table 4.

Table 4 Summary of the key requirements derived from the ALTAI framework and adapted for conceptual framework's safety-critical systems

Relevant ALTAI requirement	Relevant ALTAI sub-requirement	Conceptual framework's requirements
#1 Human agency and oversight	Human agency and autonomy	• Make sure that users are adequately informed about (1) the fact that they are interacting with an AI system and (2) the kind of inferential mechanism behind the system's output • Establish mechanisms for (1) preventing overreliance on the system and (2) monitoring the actual use of the system to constantly check for overreliance dynamics, especially in those scenarios with scarce data • Assess the risks stemming from overreliance by considering these risks in terms of hazard (the potential harming consequences of overreliance), exposure (people and assets exposed to such harm), and vulnerability • Make sure that humans maintain meaningful 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. [107])
	Human oversight	• 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

(continued)

Table 4 (continued)

Relevant ALTAI requirement	Relevant ALTAI sub-requirement	Conceptual framework's requirements
#2 Technical robustness and safety	Resilience to attacks and security	• Assess the risks stemming from potential hazards related to technical faults, outages, attacks, as well as inappropriate and malicious use • Identify the people and material assets exposed to the potential harm resulting from such hazards • Implement strategies to reduce the vulnerability to such hazards of (1) the system and (2) the exposed people and assets • Plan regular monitoring to continuously assess the involved risks and collect information on the system's real-world deployment
	General safety	• Identify possible threats by considering both their probability of occurrence and their magnitude/impact on the system • Identify the system's levels of exposure to such threats, both in terms of quantity and duration • Implement sufficient measures to make the system less vulnerable to such threats
	Accuracy	• 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
	Reliability, fallback plans, and reproducibility	• 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
#6 Societal and environmental well-being	Environmental well-being	• Identify the potential environmental impact of the system by considering both the training and the deployment phases
	Impact on work and skills	• Assess whether and how the systematic deployment of the system might cause human de-skilling by identifying: – The affected skills and the magnitude of the phenomenon – The affected workforce – 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

Table 4 (continued)

Relevant ALTAI requirement	Relevant ALTAI sub-requirement	Conceptual framework's requirements
#7 Accountability	Risk management	• Organize risk training to ensure that all three components of risk are considered • Put in place by-design mechanisms in case of applications that can adversely affect individuals in terms not only of hazard but also exposure and vulnerability
Additional requirements on risk acceptability	• 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	

6 Conclusions

This chapter presented a conceptual and technology-agnostic framework designed to integrate AI into decision-making processes in cross-sector critical infrastructures while maintaining an appropriate balance between automation and human oversight. The framework fosters collaboration between human operators and AI systems. Through an iterative co-learning approach, human operators engage with AI, refining system behavior and enhancing decision quality over time. Furthermore, the framework supports real-time operations by incorporating integrated information and predictive insights, allowing for corrective and preventive actions at different levels of automation.

More precisely, the conceptual framework was shaped by systems engineering principles, which supported the identification of requirements, functionalities, and interactions among various components. It was further grounded in three critical infrastructures—power grids, railways, and air traffic management—along with their respective use cases. Insights from human cognition and decision theory were applied to inform the decision-making model and to establish a foundation for human-AI interaction. Established procedures for requirements definition and risk assessment, such as IEC 62559-2, ISO/IEC TR 24030, and the ALTAI assessment tool, were adopted to ensure a structured and standardized approach. When integrated into the proposed framework, it provides a foundation for developers and end users to incorporate essential requirements from the initial stages of AI-based system design, fostering seamless integration and compliance with evolving regulatory landscapes.

Finally, this work highlights that integrating AI into high-risk applications and sectors extends beyond the AI component or software itself—it requires a holistic perspective that considers the entire socio-technical system in which the AI solution operates. Achieving this demands an interdisciplinary approach to shift from individual decision-making toward collaborative human-AI decision-making, drawing from fields such as philosophy and cognitive engineering. In other words, the successful adoption of emerging technologies demands a comprehensive understanding of the broader operational and social context in which they are deployed.

Acknowledgments The research leading to this work is part of the AI4REALNET (*AI for REAL-world NETwork operation*) project, which received funding from the 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). 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.

References

1. Curry, E., Heintz, F., Irgens, M., Smeulders, A. W. M., & Stramigioli, S. (2022). Partnership on AI, data, and robotics. *Communications of the ACM, 65*(4), 54–55.
2. Marot, A., Donnot, B., Chaouache, K., Kelly, A., Huang, Q., Hossain, R.-R., & Cremer, J. L. (2022). Learning to run a power network with trust. *Electric Power Systems Research, 212*, 108487.
3. Greitzer, F. L., & Podmore, R. (2008). *Naturalistic decision making in power grid operations: Implications for dispatcher training and usability testing*. Tech. rep. PNNL-18040. Pacific Northwest National Laboratory. Retrieved from https://www.pnnl.gov/main/publications/external/technical_reports/PNNL-18040.pdf
4. Fan, S., Guo, J., Ma, S., Li, L., Wang, G., Haotian, X., Yang, J., & Zhao, Z. (2024). Framework and key technologies of human machine hybrid-augmented intelligence system for large-scale power grid dispatching and control. *CSEE Journal of Power and Energy Systems, 10*(1), 1–12.
5. Hilliard, A., Brath, R., & Jamieson, G. A. (2024). Work domain analysis of electric transmission networks and operation. *IEEE Systems Journal, 18*(1), 474–484.
6. SBB. (2020). *Rail control system*. German. Retrieved April 3, 2025, from https://bahninfrastruktur.sbb.ch/de/produkte-dienstleistungen/bahninformatiksysteme/verkehrssteuerung/rcs.html
7. Rittner, M., Richta, H. N., & Große, S. (2022). *Automatische Dispositionsunterstützung mit ADA-PMB*. Retrieved October 3, 2023, from https://www.system-bahn.net/aktuell/automatische-dispositionsunterstuetzung-mit-ada-pmb/
8. Wälter, J., Mehta, F. D., & Rao, X. (2020). Aiding vehicle scheduling and rescheduling using machine learning. *International Journal of Transport Development and Integration, 4*(4), 308–320. https://doi.org/10.2495/TDI-V4-N4-308320. Retrieved March 7, 2025, from http://www.witpress.com/doi/journals/TDI-V4-N4-308-320
9. Parvez Farazi, N., Zou, B., Ahamed, T., & Barua, L. (2021). Deep reinforcement learning in transportation research: A review. *Transportation Research Interdisciplinary Perspectives, 11*, 100425. https://doi.org/10.1016/j.trip.2021.100425. Retrieved March 7, 2025, from https://linkinghub.elsevier.com/retrieve/pii/S2590198221001317
10. Li, J.-Q., Mirchandani, P. B., & Borenstein, D. (2007). The vehicle rescheduling problem: Model and algorithms. *Networks, 50*(3), 211–229. https://doi.org/10.1002/net.20199. Retrieved March 4, 2025, from https://onlinelibrary.wiley.com/doi/10.1002/net.20199
11. Mohanty, S., Nygren, E., Laurent, F., Schneider, M., Scheller, C., Bhattacharya, N., Watson, J., Egli, A., Eichenberger, C., Baumberger, C., Vienken, G., Sturm, I., Sartoretti, G., & Spigler, G. (2020). Flatland-RL: Multi-agent reinforcement learning on trains. Version number 2. https://doi.org/10.48550/ARXIV.2012.05893. Retrieved March 4, 2025, from https://arxiv.org/abs/2012.05893
12. Laurent, F., Schneider, M., Scheller, C., Watson, J., Li, J., Chen, Z., Zheng, Y., Chan, S.-H., Makhnev, K., Svidchenko, O., Egorov, V., Ivanov, D., Shpilman, A., Spirovska, E., Tanevski, O., Nikov, A., Grunder, R., Galevski, D., Mitrovski, J., Sartoretti, G., Luo, Z., Damani,

M., Bhattacharya, N., Agarwal, S., Egli, A., Nygren, E., & Mohanty, S. (2021). Flatland competition 2020: MAPF and MARL for efficient train coordination on a grid world. In H. J. Escalante & K. Hofmann (Eds.), *Proceedings of the NeurIPS 2020 Competition and Demonstration Track* (Proceedings of Machine Learning Research. PMLR) (Vol. 133, pp. 275–301). Retrieved from https://proceedings.mlr.press/v133/laurent21a.html

13. Lövétei, I., Kővári, B., Bécsi, T., & Aradi, S. (2022). Environment representations of railway infrastructure for reinforcement learning-based traffic control. *Applied Sciences, 12*(9), 4465. https://doi.org/10.3390/app12094465. Retrieved March 7, 2025, from https://www.mdpi.com/2076-3417/12/9/4465
14. Jiang, Y., Zhang, K., Li, Q., Chen, J., & Zhu, X. (2022). Multiagent path finding via tree LSTM. Version number 2. https://doi.org/10.48550/ARXIV.2210.12933. Retrieved March 7, 2025, from https://arxiv.org/abs/2210.12933
15. Roost, D., Meier, R., Huschauer, S., Nygren, E., Egli, A., Weiler, A., & Stadelmann, T. (2020). Improving sample efficiency and multi-agent communication in RL-based train rescheduling. In *2020 7th Swiss Conference on Data Science (SDS)* (pp. 63–64). IEEE. https://doi.org/10.1109/SDS49233.2020.00024. Retrieved March 7, 2025, from https://ieeexplore.ieee.org/document/9145010/
16. Shang, M., Zhou, Y., Mei, Y., Zhao, J., & Fujita, H. (2023). Energy-saving train operation synergy based on multi-agent deep reinforcement learning on spark cloud. *IEEE Transactions on Vehicular Technology, 72*(1), 214–226. https://doi.org/10.1109/TVT.2022.3205379. Retrieved March 7, 2025, from https://ieeexplore.ieee.org/document/9882357/
17. Schneider, S., Ramesh, A., Roets, A., Stirbu, C., Safaei, F., Ghriss, F., Wülfing, J., Güral, M., Siboni, N., Gentry, R., Liessner, R., Hustache, T., Lecat, T., Deekshith, U., Markin, V., Le, V., Bejjani, W., Küpper, M., & Sturm, I. (2024). Intelligent railway capacity and traffic management using multi-agent deep reinforcement learning. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), Edmonton, Canada*.
18. SNCF. (2022). Open Source Railway Designer. Retrieved April 3, 2025, from https://github.com/OpenRailAssociation/osrd
19. Nunes, T. M. M., Borst, C., van Kampen, E.-J., Hilburn, B., & Westin, C. (2021). Human-interpretable input for machine learning in tactical air traffic control. In *SESAR Innovation Days 2021*.
20. Westin, C., Hilburn, B., Borst, C., Van Kampen, E.-J., & Bång, M. (2020). Building transparent and personalized AI support in air traffic control. In *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)* (pp. 1–8).
21. Yilmaz, E., Sanni, O., Kotwicz Herniczek, M. T., & German, B. (2021). Deep reinforcement learning approach to air traffic optimization using the MuZero algorithm. In *AIAA aviation 2021 forum* (p. 2377).
22. Yutong, C., Minghua, H., Yan, X., & Lei, Y. (2023). Locally generalised multi-agent reinforcement learning for demand and capacity balancing with customised neural networks. *Chinese Journal of Aeronautics, 36*(4), 338–353.
23. Ahrenhold, N., Gerdes, I., Mühlhausen, T., & Temme, A. (2023). Validating dynamic sectorization for air traffic control due to climate sensitive areas: Designing effective air traffic control strategies. *Aerospace, 10*(5), 405.
24. Lui, G. N., Lulli, G., Lema-Esposto, M. F., & Martinez, R. L. (2024). Airspace sector design: An optimization approach.
25. 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.
26. Braunschweig, B., Gelin, R., & Terrier, F. (2022). The wall of safety for AI: Approaches in the confiance.ai program. In *Workshop on Artificial Intelligence Safety (SAFEAI).*
27. 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). Springer Nature Switzerland.

28. Dignum, Virginia. 2019. *Humane AI ethical framework. HumanE AI Deliverable 1.3*. Tech. rep. Retrieved from https://www.humane-ai.eu/wp-content/uploads/2019/11/D13-HumaneAIframework-report.pdf
29. Golpayegani, D., Pandit, H. J., & Lewis, D. (2022). Comparison and analysis of 3 key AI documents: EU's proposed AI Act, assessment list for trustworthy AI (ALTAI), and ISO/IEC 42001 AI management system. In *Artificial Intelligence and Cognitive Science (AICS 2022)* (pp. 189–200).
30. Amokrane, K., Rousseaux, V., Dussartre, M., Zouinar, M., & Renoir, N. (2024). Combining user centered design and system engineering to the design of a generic AI-based assistant. In *3rd INCOSE International Conference on Human Systems Integration (HSI), Jeju Island, South Korea*. Retrieved from https://hal.science/hal-04621899
31. Zouinar, M., Amokrane-Ferka, K., & Rousseaux, V. (2024). A user centered approach for the design of a generic AI based assistant system for work activities. In *22nd Triennial Congress of the International Ergonomics Association (IEA), Jeju Island, South Korea*. Retrieved from https://hal.science/hal-04685856
32. Roques, P. (2016). MBSE with the arcadia method and the Capella tool. In *8th European Congress on Embedded Real Time Software and Systems (ERTS 2016)*.
33. Clegg, C. W. (2000). Sociotechnical principles for system design. *Applied Ergonomics, 31*, 463–477.
34. Endsley, M. R. (2023). Supporting human-AI teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior, 140*, 107574.
35. 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.
36. National Academies of Sciences, Engineering, and Medicine. (2022). *Human-AI teaming: State-of-the-art and research needs*. The National Academies Press.
37. 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 (FAccT'23)* (pp. 333–342). Association for Computing Machinery. https://doi.org/10.1145/3593013.3594001
38. 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.
39. Ngo, T., & Krämer, N. (2022). I humanize, therefore I understand? Effects of explanations and humanization of intelligent systems on perceived and objective user understanding. PsyArXiv [Preprint]. https://doi.org/10.31234/osf.io/6az2h
40. Ha, T. W., & Kim, S. (2023). Improving trust in AI with mitigating confirmation bias: Effects of explanation type and debiasing strategy for decisionmaking with explainable AI. *International Journal of Human-Computer Interaction*, 1–12. https://doi.org/10.1080/10447318.2023.2285640
41. Endsley, M. R. (2023). Ironies of artificial intelligence. *Ergonomics, 66*(11), 1656–1668.
42. 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
43. 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
44. 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
45. 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

46. 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
47. Schaap, G., Bosse, T., & Vettehen, P. H. (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
48. Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance, 16*, 250–279. https://doi.org/10.1016/0030-5073(76)90016-7
49. Bainbridge, L. (1983). Ironies of automation. *Proceedings of IFAC, 19*(6), 775–779.
50. Polanyi, M. (2012). *Personal knowledge*. Routledge.
51. Wäfler, T., & Rack, O. (2021). Kooperation und künstliche intelligenz. In *Kooperation in der digitalen Arbeitswelt: Verlässliche Führung in Zeiten virtueller Kommunikation* (pp. 77–88).
52. Alexander, P. A., Schallert, D. L., & Reynolds, V. E. (2009). What is learning anyway? A topographical perspective considered. *Educational Psychologist, 44*(3), 176–192.
53. Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice-Hall.
54. Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. arXiv preprint arXiv:2010.07487. Retrieved from http://arxiv.org/abs/2010.07487
55. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*, 50.
56. 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
57. 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
58. Koopman, P., & Hoffman, R. R. (2003). Work-arounds, make-work, and kludges. *IEEE Intelligent Systems, 18*(6), 70–75.
59. 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
60. Behzadan, V., & Munir, A. (2017). Whatever does not kill deep reinforcement learning, makes it stronger. arXiv preprint arXiv:1712.09344.
61. Zissis, G. (2019). The r3 concept: Reliability, robustness, and resilience [president's message]. *IEEE Industry Applications Magazine, 25*(4), 5–6.
62. Molnar, C. (2020). *Interpretable machine learning*. Lulu.com. Retrieved from https://christophmolnar.com/books/interpretable-machine-learning/
63. Vouros, G. A. (2022). Explainable deep reinforcement learning: State of the art and challenges. *ACM Computing Surveys, 55*, 1–39.
64. European Commission (EC). (2024). *Ethics guidelines for trustworthy AI*. Tech. rep. Retrieved from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
65. Nichol, A., Pfau, V., & Hesse, C. (2018). Gotta learn fast: A new benchmark for generalization in RL. arXiv preprint arXiv:1804.03720. https://arxiv.org/abs/1804.03720
66. Irpan, A. (2018). *Deep reinforcement learning doesn't work yet*. Retrieved February 15, 2025, from https://www. alexirpan.com/2018/02/14/rl-hard.html
67. Cobbe, K., Klimov, O., & Hesse, C. (2019). Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning* (pp. 1282–1289).
68. Olteanu, A., Castillo, C., & Diaz, F. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data, 2*, 13. https://doi.org/10.3389/fdata.2019.00013

69. Paliouras, G. (1993). *Scalability of machine learning algorithms*. Doctoral dissertation, University of Manchester.
70. Ulanov, A., Simanovsky, A., & Marwah, M. (2017). Modeling scalability of distributed machine learning. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)* (pp. 1249–1254).
71. 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. https://doi.org/10.1007/s10458-019-09421-1
72. Cobb, A. D., Jalaian, B., Bastian, N. D., & Russell, S. (2021). Toward safe decision-making via uncertainty quantification in machine learning. In *Systems engineering and artificial intelligence* (pp. 379–399).
73. Nemani, V., Biggio, L., Huan, X., Hu, Z., Fink, O., Tran, A., Wang, Y., Zhang, X., & Hu, C. (2023). Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial. *Mechanical Systems and Signal Processing, 205*, 110796.
74. 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.
75. Von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., et al. (2021). Informed machine learning—A taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering, 35*(1), 614–633.
76. Araki, B., Li, X., Vodrahalli, K., DeCastro, J., Fry, M., & Rus, D. (2021). The logical options framework. In *International Conference on Machine Learning* (pp. 307–317).
77. Lyu, D., Yang, F., Liu, B., & Gustafson, S. (2019). SDRL: Interpretable and data-efficient deep reinforcement learning leveraging symbolic planning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 2970–2977).
78. Vaezipoor, P., Li, A. C., Icarte, R. A. T., & McIlraith, S. A. (2021). LTL2action: Generalizing LTL instructions for multi-task RL. In *International Conference on Machine Learning* (pp. 10497–10508).
79. Yang, F., Lyu, D., Liu, B., & Gustafson, S. (2018). PEORL: Integrating symbolic planning and hierarchical reinforcement learning for robust decision-making. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (pp. 4860–4866). AAAI Press.
80. Van der Pol, E., Worrall, D., van Hoof, H., Oliehoek, F., & Welling, M. (2020). MDP homomorphic networks: Group symmetries in reinforcement learning. *Advances in Neural Information Processing Systems, 33*, 4199–4210.
81. Höpner, N., Tiddi, I., & van Hoof, H. (2022). Leveraging class abstraction for commonsense reinforcement learning via residual policy gradient methods. In *International Joint Conference on Artificial Intelligence, IJCAI 2022* (pp. 3050–3056).
82. Gao, J., Chen, S., Li, X., & Zhang, J. (2022). Transient voltage control based on physics-informed reinforcement learning. *IEEE Journal of Radio Frequency Identification, 6*, 905–910.
83. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in neural information processing systems* (p. 30).
84. Kaplan, R., Sauer, C., & Sosa, A. (2017). Beating atari with natural language guided reinforcement learning. arXiv preprint arXiv:1704.05539.
85. Knox, W. B., & Stone, P. (2009). Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the International Conference on Knowledge Capture* (pp. 9–16).
86. Charpentier, B., Senanayake, R., Kochenderfer, M., & Günnemann, S. (2022). Disentangling epistemic and aleatoric uncertainty in reinforcement learning. arXiv preprint arXiv:2206.01558.

87. Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning, 110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3
88. Palminteri, S., & Lebreton, M. (2021). Context-dependent outcome encoding in human reinforcement learning. *Current Opinion in Behavioral Sciences, 41*, 144–151. https://doi.org/10.1016/j.cobeha.2021.06.012
89. Nylin, M., Westberg, J. J., & Lundberg, J. (2022). Reduced autonomy workspace (raw)—An interaction design approach for human automation cooperation. *Cognition, Technology & Work, 24*(2), 261–273. https://doi.org/10.1007/s10111-021-00674-1
90. Bondi, E., Koster, R., Sheahan, H., Chadwick, M., Bachrach, Y., Cemgil, A. T., et al. (2022). Role of human-AI interaction in selective prediction. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*, 5286–5294.
91. 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* (pp. 572–589).
92. 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). https://doi.org/10.1007/s10458-022-09564-8
93. 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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11760769
94. 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.
95. 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
96. 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. https://doi.org/10.1016/j.ssci.2005.10.019
97. Marot, A., Kelly, A., Naglic, M., Barbesant, V., Cremer, J., Stefanov, A., & Viebahn, J. (2022). Perspectives on future power system control centers for energy transition. *Journal of Modern Power Systems and Clean Energy, 10*(2), 328–344. https://doi.org/10.35833/MPCE.2021.000687
98. Amokrane-Ferka, K., Marot, A., Meddeb, M., Dussartre, M., Crochepierre, L., Rozier, A., Renoir, N., Gosselin, S., Girod, H., Khouadjia, M., et al. (2024). Framework for human and AI assistant bidirectional interaction applied to industrial system operations.
99. Metzinger, T. (2019). Ethics washing made in Europe. *Der Tagesspiegel*. https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html
100. Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*, 2749.
101. Zanotti, G., Petrolo, M., Chiffi, D., & Schiaffonati, V. (2024). Keep trusting! A plea for the notion of trustworthy AI. *AI & Society, 39*(6), 2691–2702.
102. Nickel, P. J., & Vaesen, K. (2012). Risk and trust. In S. Roeser, R. Hillerbrand, M. Peterson, & P. Sandin (Eds.), *Handbook of risk theory* (pp. 857–870). Springer. https://doi.org/10.1007/978-94-007-1433-5_34
103. Royal Society (Great Britain). (1983). *Risk assessment: Report of a Royal Society study group*. Royal Society.

104. Hansson, S. O. (2009). From the casino to the jungle: Dealing with uncertainty in technological risk management. *Synthese, 168*, 423–432. https://doi.org/10.1007/s11229008-9447-0

105. Nordström, M. (2022). Ai under great uncertainty: Implications and decision strategies for public policy. *AI & Society, 37*, 1703–1714. https://doi.org/10.1007/s00146022-01393-1

106. van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics, 22*, 667–686.

107. Prunkl, C. (2022). Human autonomy in the age of artificial intelligence. *Nature Machine Intelligence, 4*, 99.