AI for real-world network operation

WP5 – Dissemination, communication, and exploitation of results

D5.5 – Communication and dissemination monitoring phase 2

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

DOCUMENT INFORMATION

DOCUMENT	D5.5 – Communication and dissemination monitoring phase 2
TYPE	Report
DISTRIBUTION LEVEL	Public
DUE DELIVERY DATE	31/03/2026
DATE OF DELIVERY	25/03/2026
VERSION	V1.0
DELIVERABLE RESPONSIBLE	INESC TEC
AUTHOR (S)	Catarina Carvalho (INESC TEC), Alípio Torre (INESC TEC)
OFFICIAL REVIEWER/s	Ricardo Bessa (INESC TEC), Milad LEYLI-ABADI (IRTSX)

DOCUMENT HISTORY

VERSION	AUTHORS	DATE	CONTENT AND CHANGES
0.1	Catarina Carvalho, Alípio Torre	06/02/2026	First draft version
0.2	Ricardo Bessa	06/02/2026	Revision
0.3	Catarina Carvalho, Alípio Torre	03/03/2026	Revision
0.4	Milad LEYLI-ABADI	11/03/2026	Review
0.5	Catarina Carvalho, Alípio Torre	20/03/2026	Final revision
1.0	Ricardo Bessa	25/03/2026	Final version

ACKNOWLEDGEMENTS

NAME	PARTNER
Ricardo Bessa	INESC TEC

DISCLAIMER

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

EXECUTIVE SUMMARY

This report presents the fifth deliverable to be submitted within the scope of the AI4REALNET project work package 5 (WP5): "Communication and dissemination monitoring phase 2".

The project's WP5 is structured in four tasks, which will be developed and implemented throughout the whole project, resulting in a total of nine deliverables [1]:

- Task 5.1 - Dissemination and communication plan
- Task 5.2 - Dissemination boosters
- Task 5.3 - Cooperation and synergies with regional and European initiatives/stakeholders
- Task 5.4 - Exploitation strategy and Plan

This deliverable, "Communication and dissemination monitoring phase 2", intends to provide an overview of the work developed between months 19-30 of the project, regarding the communication and dissemination activities, established in the D5.1 - "Communication and Dissemination Plan" [2]. Therefore, this report is divided into two main parts: the communication and the dissemination plans.

On one hand, the dissemination part is focused on a macro strategy, with the presentation of the KPIs achieved within the communication actions carried out. On another, the communication part monitors the results of the integrated marketing approach based on several communication activities.

Even though all results revealed will be detailed further in the document, an overview is first presented.

The measurable WP5 achievements about dissemination and communication during this reporting period include: 1 project's website with a total of 4.641 visits; 3 dedicated social media channels continuously operating (883 followers on LinkedIn, 87 followers on X and 29 subscribers on YouTube); 4 videos produced, including 2 new technical videos approaching the project's results, all summing a total of 119 views on YouTube; 2 brochures; 3 newsletters, that reached a total of 194 subscribers; and 2 articles published on other online platforms, one of them mentioning the project as one of the EU preparedness union strategy highlight programme.

The dissemination efforts included a total of 33 events: 7 organised by the consortium and 26 in which the partners participated. The events organised included two workshops at the ECML PKDD 2025 conference, one special session at the IHIET International Conference 2025, and a Hackathon dedicated to the railway area.

The Advisory Board, composed of 20 expert members from diverse sectors and regions, was formally established, and two meetings were held – one in December 2024 and another in July 2025. Their feedback has been instrumental in refining evaluation protocols, strengthening the exploitation strategy, and ensuring that AI4REALNET's solutions effectively address real-world needs across critical infrastructure sectors.

AI4REALNET disclosed a GitHub repository and page created for disseminating the project's results in public and free access, as well as released information through the AI-on-demand platform (AIoD). Regarding the publications on GitHub, these summed a total of 217 views and 161 contributors. A dedicated Zenodo community was established, and, in the first 30 months, 18 items were uploaded, collectively reaching 1.220 views and 1.108 downloads, with the data set developed in 2024 taking the lead both in views and downloads.

13 publications were issued, all available in open access, and 1 remains under review: among these, 3 publications have been submitted and presented at IEEE Power Tech 2025, 5 at the ECML PKDD 2025 Conference, and 2 at the IHIET 2025 Conference. This commitment to promote the work developed follows the results presented on the previous monitoring deliverable (D5.3), where a total of 15 scientific publications were presented, of which 4 were submitted and presented in the ECML 2024 Machine Learning for Sustainable Power Systems (ML4SPS) workshop, and 3 in the Annual Conference on Neural Information Processing Systems (NeurIPS 2024).

A collaboration between projects from the HORIZON-CL4-2022-HUMAN-02-01 call is ongoing, and that includes TANGO, THEMIS 5.0, PEER, and HumAIne projects. This synergy resulted in a joint policy brief entitled "AI Your Way – Trusted, Scalable & Ready to Deliver", as well as the organisation of a webinar in January 2026 - "From Command to Collaboration: Reimaging the human-AI relationship AI, Data and Robotics Forum".

Also, the cooperation with the Adra-e project is a key approach to the project, aiming to create the conditions for an inclusive and sustainable European ecosystem in AI, Data and Robotics. Within the multiple joint activities, highlights include participation in the ADRF 2025 event and the contribution to the ADRA Strategic Research, Innovation, and Deployment Agenda (SRIDA). Two book chapters have been submitted to the Adra-e open-access book "Artificial Intelligence, Data and Robotics: Foundations, Transformations, and Future Directions" and were accepted for publication.

AI4REALNET's commitment to advancing the European AI ecosystem extends beyond its core technical objectives through strategic partnerships and contributions to major collaborative initiatives. The project has actively participated in ADRA-e Strategic Research, Innovation, and Deployment Agenda

(SRIDA), with the project coordinator chairing the Energy Topic Group and providing critical input on AI deployment in critical energy infrastructures. Two peer-reviewed book chapters have been accepted for publication in ADRA-e's open-access book by Springer, addressing human-AI collaboration frameworks and visualization perspectives. Beyond ADRA-e, AI4REALNET has also established strategic cooperation with the Linux Foundation Europe to ensure long-term sustainability of key digital environments, with Grid2Op successfully integrated as an open-source project within LF Energy. Additionally, memoranda of understanding with projects such as ELOQUENCE, combined with active engagement through platforms like AI-on-Demand and contributions to the ADR Awareness Centre, demonstrate AI4REALNET's role as a catalyst for collaborative innovation and knowledge exchange across Europe's AI and critical infrastructure communities.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

AB	Advisory Board
AI	Artificial Intelligence
AIoD	AI on Demand
ATM	Air Traffic Management
CRM	Customer Relationship Management
D	Deliverable
EC	European Commission
EU	European Union
GA	Grant Agreement
GDPR	EU General Data Protection Regulation
IMC	Integrated Marketing Communication
IMS	Integrated Marketing Strategy
KPI	Key Performance Indicator
LF	Linux Foundation
M	Month
MUAC	Maastricht Upper Area Control Centre
N/A	Not available
SME	Small and Medium-sized Enterprise
TEF	Testing and Experimentation Facility
UC	Use Case
WP	Work Package

1. INTRODUCTION

The WP5 “Dissemination, communication, and exploitation of results” of the AI4REALNET project has the following objectives:

- Deliver communication, dissemination, engagement, and cooperation strategies.
- Deliver relevant input to scientific communities.
- Promote open science activities.
- Ensure collaboration with AI4Europe and Adra-e projects.
- Benchmark and engage the AI community along the project and get contributions (e.g., participation in open innovation with the AI4REALNET digital environments).
- Promote the exploitation of the AI4REALNET results and of the technologies.

The first deliverable, D5.1 – Communication and dissemination plan [2], sets up a dissemination and communication plan, following an Integrated Marketing Strategy approach (IMS), to reach certain outcomes from the beginning until the end of the project.

The third WP5 deliverable, D5.3 – Communication and dissemination monitoring phase 1 [3], analysed the work developed within the WP, providing an overview of the outcomes achieved in the first 18 months of the project.

In this deliverable, a round of monitoring of the communication and dissemination activities of the project takes place, following the previous analysis and, consequently, assessing the course of work developed. The project dissemination chapter will present the outcomes by communication tool, as outlined in D5.1, and conclude with a summary of the KPIs achieved so far. Regarding the project communication chapter, it will present the results of the third communication campaign and outline the next steps, thereby approaching the fourth and final campaign.

2. PROJECT DISSEMINATION

The AI4REALNET project dissemination plan (D5.1) outlined a set of strategic, audience-focused actions to promote the project and its results, along with measurement tools and indicators to evaluate the impact of each activity. This report describes the outcomes achieved during months 19 to 30 of the project, building on D5.3, which covered dissemination results from the first 18 months of implementation.

2.1 DISSEMINATION IMPACT ASSESSMENT (MONTHS 19-30)

According to the categorisation of the communication tools in deliverable D5.1, the results achieved so far for each of the actions developed will be presented in seven subsections: Advertising, Digital Marketing, Direct Marketing, Public Relations, Advisory Board, Open Science and Cooperation with Stakeholders.

Before presenting in detail the outcomes achieved, a general overview is presented in Figure 1. The details about each outcome are presented in the following subsections. The indicators have been recorded by the uOneConnect platform, the project's management tool that consists of a web-based application that addresses all the work developed throughout the project (more information about this platform can be found in deliverable D6.1, "Project management guide – procedures handbook" [4]).

The main outcomes per category between months 19-30 in absolute results are the following:

- Social media channels: 160 posts
- Newsletters: 3 sent
- Articles: 2
- Events: 33
- Scientific Publications issued: 13
- Scientific Publications in review: 1

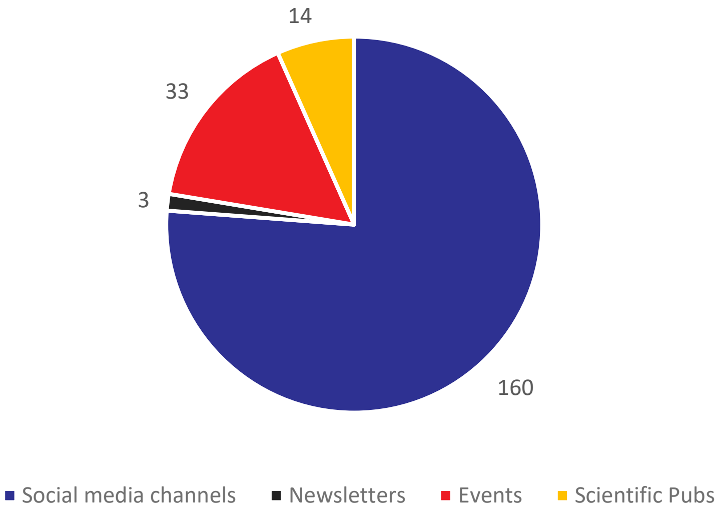


FIGURE 1 - MAIN OUTCOMES PER CATEGORY OF THE PROJECT

During the period from months M19 to M30, the project carried out a range of actions to assess dissemination impact. Based on the categories outlined in the results above, Figure 2 illustrates the different types of actions undertaken and their relative weight within the overall communication activity. As shown, social media engagement represents the most significant share, emerging as the most relevant channel for reaching the project's target audiences to date. This is followed by events, which also demonstrate the notable impact of this tool for engaging with stakeholders of interest. This stands out from the results in the first monitoring report, which showed that the news publishing was the most significant dissemination tool. As the project evolved, the participation and organisation of events naturally emerged as the most valuable tool for reaching dedicated audiences.

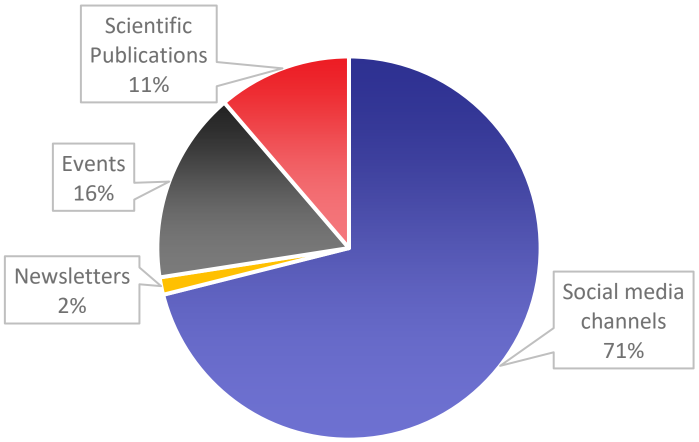


FIGURE 2 – MAIN COMMUNICATION OUTCOMES PER CATEGORY IN PERCENTAGE

2.1.1 ADVERTISING

Under the Integrated Marketing Communication approach – the marketing strategy adopted for the project – a set of marketing tools was chosen to support the communication process, each one fulfilling a distinct function. Collectively, these tools are intended to support engagement with the project's target audiences and to contribute to achieving the project's objectives.

So, activities related to tools such as advertising, namely throughout the design of the project identity, presentation, x-banner, and slide deck, were already developed and presented in the deliverable D5.1. The next subsections will focus on the work regarding the advertising strategy that followed D5.1, and that extends to the end of the project, namely brochures and videos. In the previous report, a set of infographics was mentioned in this chapter as a relevant communication tool. However, during the reporting period in question, no need was found to support the development of such design materials.

2.1.1.1 BROCHURES

Between months 19 and 30, two brochures were planned and executed (**ANNEX 1**), one regarding the project's dissemination and the other developed specially for an event participation. From the first one, 50 printed copies were produced and delivered, as well as published on the project's website. This digital version was made available to the entire consortium, one for digital use only and the other specifically for printing. Regarding the second brochure, 100 printed copies were produced and delivered during the event in question (Web Summit 2025).

2.1.1.2 VIDEOS

The initial communication plan included the production of three videos targeting diverse project audiences. During the reporting period, four additional videos have been produced and disseminated, bringing the total to 26 videos has already been produced, including 12 interviews, a project video presentation, the recording and editing of four webinars, and four technical videos regarding the work developed. All videos are available on the project's YouTube page. One of the videos produced is unlisted (available through a targeted link) because it served as a project presentation for an online event. Table 1 presents an overview of the four videos produced during months 19 and 30, and respective views, registered on March 25, 2026.

Title	Date	Views
AI4REALNET Seasons Greetings 2025	December 21, 2025	--
AI Assistant with a Hypervision-Based HMI for Power Grids	November 13, 2025	84

Title	Date	Views
BlueSky plugin for deploying reinforcement learning models	September 25, 2025	22
AI4REALNET WP5 presentation @ project's 3rd General Assembly	June 17, 2025	5

TABLE 1 - LIST OF THE VIDEOS PRODUCED BY THE PROJECT IN M19-30

2.1.2 DIGITAL MARKETING

This subsection outlines the results achieved during the project's 12 months between M19-30, focusing on the website and social media channels.

2.1.2.1 WEBSITE

The AI4REALNET project website was launched at the end of January 2024. It started with a related measuring tool in operation, namely the Google Analytics platform. However, due to the EU General Data Protection Regulation (GDPR), it was replaced with the Matomo platform.

Until March 25, 2026, the website had 6.530 visits and totalled 17.785 page views. Table 2 - Top 5 pages visited on the project website shows the five most visited pages of the website – highlights for the homepage with a great distance from other individual pages, but followed by the project details page –, and Table 3 presents the top five visitors' origin countries – with United States visitors leading the table, followed by Portugal, Switzerland, Germany, and Italy, numbers justified by IP registrations. The map in Figure 3 – Worlwide map with top countries of origin illustrates the countries distribution worldwide.

Page title	Page views
Homepage	4.317
Project	1.112
Use Cases portfolio	1.074
Use Cases homepage	931
Scientific Papers	736

TABLE 2 - TOP 5 PAGES VISITED ON THE PROJECT WEBSITE

Country	Visits
United States	1.538
Portugal	942
Switzerland	578
Germany	559
Italy	315

TABLE 3 - TOP 5 COUNTRIES OF ORIGIN

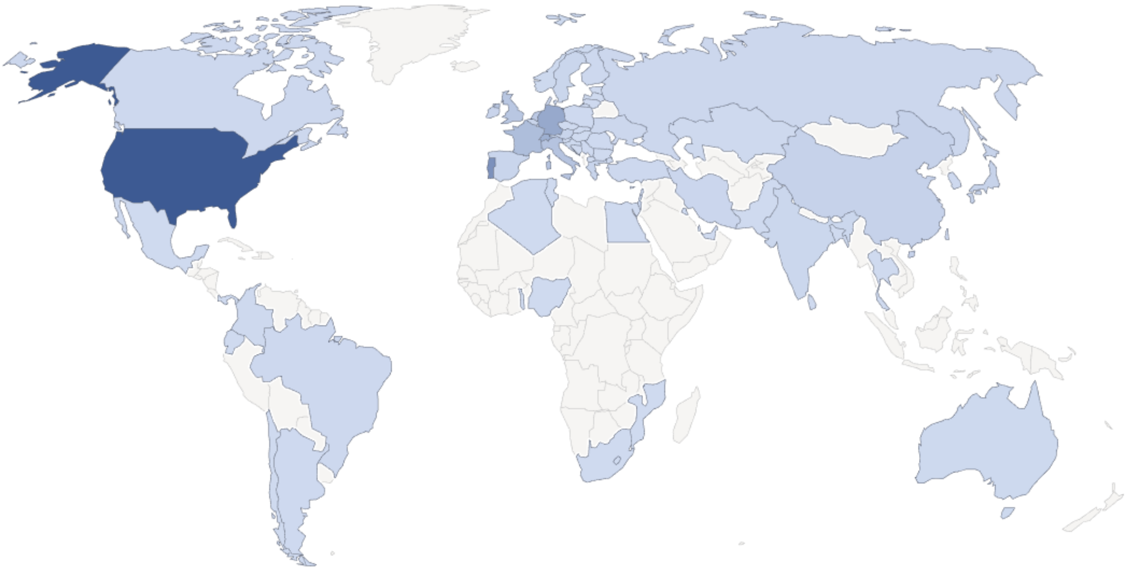
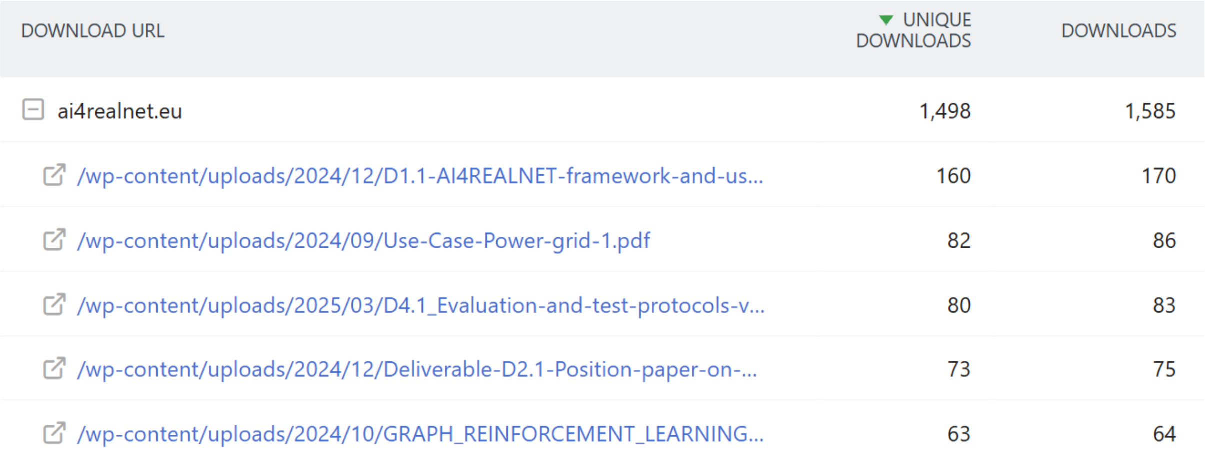


FIGURE 3 – WORLWIDE MAP WITH TOP COUNTRIES OF ORIGIN

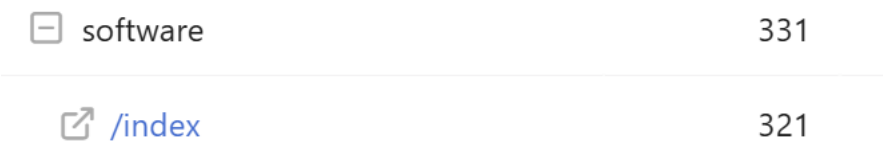
Regarding the website's document downloads, from a total of 1.585 downloads, the deliverables are in the top five of the list, together with the use cases. The next figure shows the most downloaded documents from the website.



DOWNLOAD URL	▼ UNIQUE DOWNLOADS	DOWNLOADS
⊟ ai4realnet.eu	1,498	1,585
/wp-content/uploads/2024/12/D1.1-AI4REALNET-framework-and-us...	160	170
/wp-content/uploads/2024/09/Use-Case-Power-grid-1.pdf	82	86
/wp-content/uploads/2025/03/D4.1_Evaluation-and-test-protocols-v...	80	83
/wp-content/uploads/2024/12/Deliverable-D2.1-Position-paper-on-...	73	75
/wp-content/uploads/2024/10/GRAPH_REINFORCEMENT_LEARNING...	63	64

FIGURE 4 – TOP 5 DOWNLOADED DOCUMENTS

It is worth noting that the software available in open source through the project's website is not directly downloadable, meaning that the user must follow a link to access the GitHub page. So, there are no indicators on how many times the software was reached through the website; the information that can be collected is only regarding the views of that page, with a total of 331, shown in Figure 5 – Views of the website's software page.



⊟ software	331
/index	321

FIGURE 5 – VIEWS OF THE WEBSITE'S SOFTWARE PAGE

2.1.2.2 SOCIAL MEDIA

Since the beginning of the project and until March 25, 2026, 386 posts were made on all the project's social media channels – LinkedIn, X (former Twitter), and YouTube.

Figure 6 - Percentage of posts done on each of the social media channels presents the percentage of posts done on each of the project's social media channels during the whole period of the project. The absolute numbers are:

- LinkedIn: 208 posts
- X: 163 posts
- YouTube: 24 posts

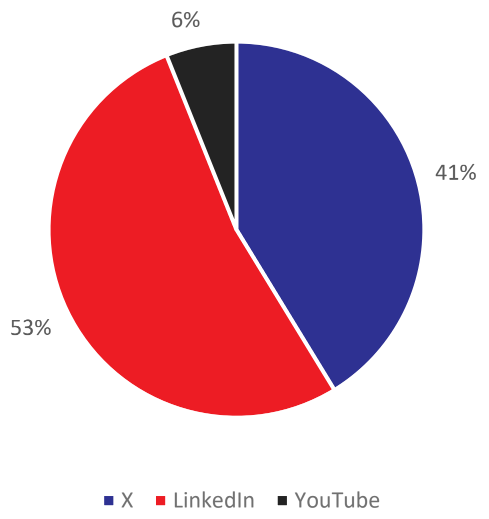


FIGURE 6 - PERCENTAGE OF POSTS DONE ON EACH OF THE SOCIAL MEDIA CHANNELS

Considering the reporting period in question (M19-30), there was a 37% increase in communication across all social media pages. The absolute numbers for the posts done in this period are described below:

- LinkedIn: 85 posts
- X: 71 posts
- YouTube: 4 posts

Keeping in mind that each social media platform has its own algorithm, here are some key points to keep in mind for this strategy during the reporting period in question.

Regarding LinkedIn, the AI4REALNET profile currently has 883 followers. In the reporting period in question, the page had a total of 44.400 organic impressions (meaning the number of times the content has naturally appeared on someone's page), 980 reactions, 12 comments, and 23 reposts to our publications. Analysing the posts made between months 19 and 30, the highlight is information about scientific papers, followed by workshops and events. The interaction indicators show that the following are the five posts with more engagement:

Post Title	Date	Type	Impressions	Clicks	Engagement Rate
And that's a wrap! The AI4REALNET Project 3rd assembly in Politecnico de Milano has come to an end	June 4, 2025	Image	4.574	535	14,25%
"Machine Learning for Sustainable Power Systems" Workshop Call for papers open	April 24, 2025	Image	2.101	94	7,09%
[Reminder] Workshop "Machine Learning for Sustainable Power Systems" Call for papers OPEN	May 22, 2025	Repost	1.969	64	5,38%
[New research article in open access] "Centrally Coordinated Multi-Agent Reinforcement Learning for Power Grid Topology Control"	June 23, 2025	Article	1.255	58	6,29%
New paper published: "Learning Topology Actions for Power Grid Control"	October 21, 2025	Image	1.196	71	7,86%

TABLE 4 – TOP 5 LINKEDIN POSTS BY NUMBER OF IMPRESSIONS

Regarding the X's project profile, it has 87 followers, and during the reporting period in question, it collected 4.442 impressions (the number of times a post has been seen on users' screens), 168 likes, and 53 reposts. The changes to the platform's algorithm continue to affect this social media network's impact, as shown by the results presented, making it difficult to counteract. Even though, the most engaging posts are related to participation in events, namely the Web Summit 2025, as shown below.

Title	Date	Impressions	Engagements
"AI your way – Trusted, Scalable & Ready to Deliver" Policy brief	Oct 30, 2025	68	5
AI4REALNET at the LFE Foundation Summit Europe 2025	Jul 25, 2025	63	5
AI4REALNET Project starting at stand E326 WebSummit 2025	Nov 11, 2025	60	4
118 speakers at the LF Energy Summit Europe 2025	Aug 27, 2025	53	5

Don’t miss our masterclass at WebSummit 2025	Nov 11, 2025	51	5

TABLE 5 – TOP FIVE X POSTS BY NUMBER OF IMPRESSIONS

The YouTube page has 29 subscribers at this time and had a total of 1.448 views. The fact that the videos are constantly online and disseminated makes it difficult to determine the number of views for specific videos over a defined period. However, it has hosted three more videos on the project’s outcomes since month 19, totalling 119 views. The individual impact of each of these videos was already presented in Chapter 2.1.1.2 Videos. Since all the videos available on YouTube are also shared on other social media channels, this effect might explain the channel’s lower reach compared to others. Nevertheless, the following picture presents the five most seen videos during the reporting period, with a scientific result demo taking the lead.

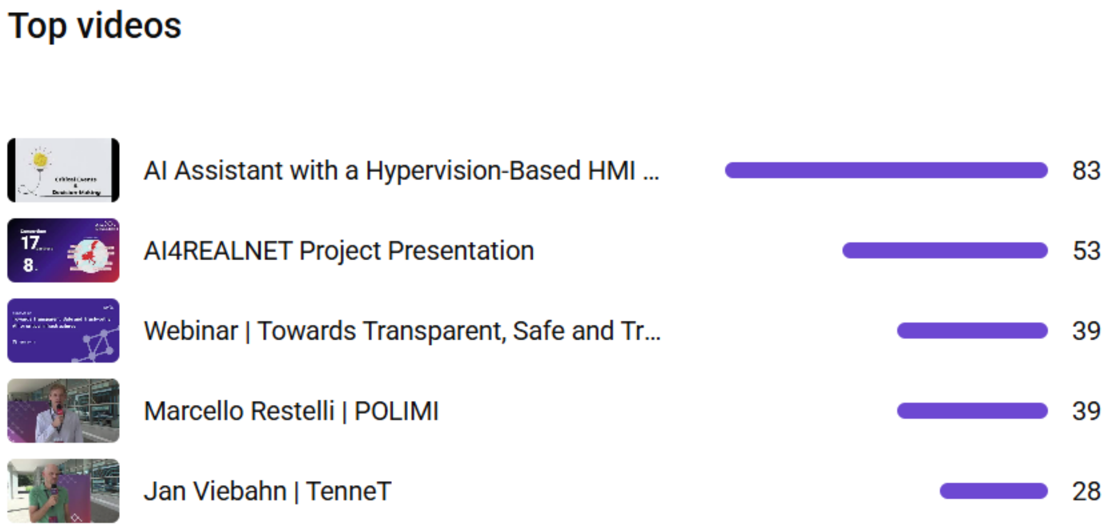


FIGURE 7 – TOP 5 VIDEOS OF THE PROJECT’S YOUTUBE PAGE

2.1.3 DIRECT MARKETING

Direct marketing involves using tools to interact directly with a defined target audience. Within this project, particular emphasis is placed on the use of newsletters, as outlined below.

2.1.3.1 NEWSLETTERS

Since the beginning of the project, seven newsletters have been produced and disseminated, both launched by e-mail and published on the project’s website (**ANNEX 2**). These are dedicated specially to the project and its technical content information. In parallel, three more newsletters were developed, but this time dedicated to the dissemination of a specific event (the AI for Safety-Critical Infrastructures (AI-SCI) workshop at the ECML PKDD 2025 conference). Hence, during the reporting period in question, and from those 10 newsletters, the communication team delivered three institutional newsletters and three more dedicated to an event.

The mailing has been implemented through a CRM (Customer Relationship Management) platform, namely Mailchimp, from which it is possible to monitor the reach of each posting. Audiences are challenged to subscribe to the newsletter through the website, in several events organised by the consortium, and in calls to action for that subscription regularly published on social media channels. Until the date of this report, 194 contacts have been added to the project's audience.

All project newsletters launched between M19-30 can be consulted below:

Title	Date	Recipients	Opening rate
AI4REALNET Newsletter #5	April 30, 2025	183	35.4%
AI4REALNET Newsletter #6	September 30, 2025	189	49.5%
AI4REALNET Newsletter #7	January 30, 2026	189	42.5%

TABLE 6 - PROJECT'S LAUNCHED NEWSLETTERS BETWEEN MONTHS 19-30

2.1.4 PUBLIC RELATIONS

According to D5.1, in this subsection, four actions regarding the project's communication strategy are to be introduced: press releases, news pieces published in the media, news pieces published on other platforms, and events. That was clear on the D5.3 report, where the outcomes of the first three instruments were much more evident than the latter, the events, and that was crucial to the objectives of the two initial communication campaigns: deliver strategies to create awareness of the project's activities and engage with all audiences.

However, in the reporting period in question, the reverse happens, since the purpose is now to reinforce the technical messages regarding the project's results, targeting relevant but specific audiences. So, the events take a leading role, instead of the rest of the actions mentioned, and that is the focus of the following chapters.

2.1.4.1 NEWS PIECES PUBLISHED ON OTHER PLATFORMS

Between M19-30, one opinion article was published in IEEE Spectrum, a technology magazine dedicated to engineering and the applied sciences, addressing the work developed in the AI4REALNET project in the scope of the Iberian energy blackout that happened in April 2025. In addition, the European Health and Digital Executive Agency (HaDEA) launched an article focusing on the EU preparedness union strategy, in which the AI4REALNET project is one of the highlighted EU funding programmes that strengthen the Union's resilience and capacity to respond to any future crisis. Table 7 - Articles published mentioning the project lists the articles' detailed information.

Title	Date	Venue	Type
Rules, Not Renewables, Might Explain the Iberian Blackout	June 17, 2025	IEEE Spectrum	Article
EU Preparedness – Projects reinforcing Europe's resilience in a changing world	January 16, 2026	EC HaDEA	Article

TABLE 7 - ARTICLES PUBLISHED MENTIONING THE PROJECT

2.1.4.2 EVENTS

Two types of events can occur within the AI4REALNET project: organised by the consortium or by other entities or institutions (conferences, EU events, workshops, etc.) in which the project's partners participate. The percentage of those events can be observed in Figure 8 - Percentage of Events regarding the project's Consortium.

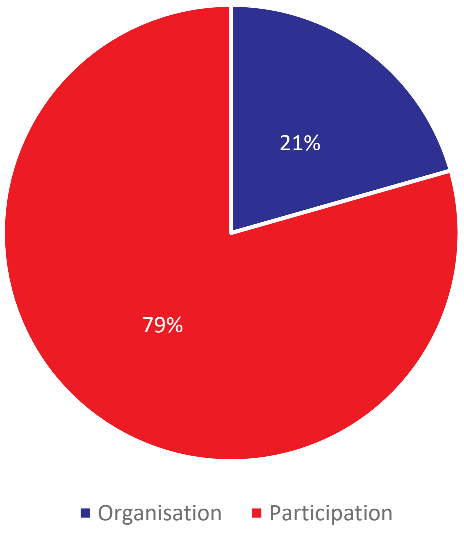


FIGURE 8 - PERCENTAGE OF EVENTS REGARDING THE PROJECT'S CONSORTIUM

Regarding the consortium's events, 6 events occurred within M19-30, namely two workshops and one special session, with an average of 150 participants.

Title	Date	Partners	Audience
Workshop "AI for Safety-Critical Infrastructures"' @ECMLPKDD 2025 – see a summary of conclusions of this workshop in ANNEX 3 **– ECML PKDD 2025 workshop on critical infrastructures** The set of papers presented at the workshop shows that trustworthy AI for critical infrastructures is heading away from isolated algorithmic performance claims and towards system-level assurance, where safety, robustness, accountability, and operational feasibility must be engineered together. Across the contributions, a recurring theme emerges: critical infrastructures cannot afford to treat AI as a "smart component" inserted into an otherwise deterministic control chain. Instead, AI must be embedded as a managed and supervised element of a socio-technical system, continuously evaluated, constrained, and aligned with safety envelopes and human operational priorities. A central contribution of the workshop papers is the explicit recognition that traditional correctness is insufficient for autonomous or AI-driven decision-making in critical infrastructures. Atif et al. argue that complex ML-based functions inherently face "unknown inputs" that cannot be exhaustively tested, thus undermining classical verification approaches and certification logic. Their position reframes the problem: infrastructures should not demand perfect correctness, but rather trustworthiness, supported through architectural patterns such as monitoring, redundancy, rejection mechanisms, and diversity (Atif et al., 2025). This is particularly relevant for infrastructures such as transportation networks, energy systems, or automated surveillance, where operational environments are open-ended, and the cost of failure is high. Their argument introduces a	September 15, 2025	INESC TEC, FLATLAND, UvA, POLIMI	25 participants

Title	Date	Partners	Audience
key challenge: infrastructures need engineering methods to make systems dependable even when components behave unpredictably, rather than attempting to eliminate unpredictability entirely (Atif et al., 2025). This is a major shift in how safety cases for AI-enabled infrastructure control can be constructed. A second important contribution lies in the reinforcement learning (RL) domain, where several papers highlight both the promise and the fragility of RL for infrastructure control. King et al. explore a cruise ship HVAC optimization scenario in which RL is used to improve energy efficiency under uncertain environmental and operational conditions. Their work introduces an important link between classical hazard analysis and RL safety: they use System-Theoretic Process Analysis (STPA) to identify hazards and translate them into constraints and controller requirements, then implement these constraints and requirements as a safety shield that blocks unsafe actions (King et al., 2025). The study reveals a practical challenge: shielding improves safety during deployment, but blocking unsafe actions during training can hinder the RL agent's ability to learn a safe policy. This reflects a critical infrastructure dilemma: safety enforcement mechanisms may distort learning dynamics, creating a tension between exploration and constraint satisfaction that is rarely acknowledged in theoretical RL research. Their paper thus contributes an applied insight: safety must be treated not only as a deployment property, but also as a training-time design variable, particularly when infrastructures demand predictable behavior from the first day of operation (King et al., 2025). This challenge is further reinforced by the broader perspective provided by (Yamagata et al., 2025), which addresses the fragmentation of terminology and methodology in safe RL. The authors propose			

Title	Date	Partners	Audience
consensus definitions and a structured checklist spanning ethical requirements, reward specification, robustness against uncertainty, explicit constraints, the use of simulators and expert knowledge, and finally, safety layers such as shielding and human intervention. For critical infrastructures, this paper's key contribution is not a new algorithm but rather a structured engineering viewpoint: safe RL cannot be reduced to constrained optimization alone. It must incorporate traceability, explainability, reward alignment, and mechanisms for human override. The paper also highlights a major infrastructure challenge: the lack of standardized guidelines makes it difficult for operators and regulators to judge whether an RL-based controller is deployable. In safety-critical domains, such ambiguity is itself a risk, because it blocks reproducibility and certification readiness. The workshop also provides evidence that AI safety challenges in infrastructures are increasingly tied to security and adversarial resilience, not only physical failures. Lowe et al. contribute a strong example in the space domain, focusing on satellite object detection. Their work proposes a statistical method based on Mahalanobis distance to distinguish between adversarial perturbations and natural sensory anomalies. Their findings show that adversarial attacks can degrade detection performance at perturbation levels where natural noise does not trigger detection, and that common anomaly detection methods such as variational autoencoders may fail to detect adversarial manipulations (Lowe et al., 2025). This contribution is highly relevant to critical infrastructure because it highlights the necessity of threat differentiation: an infrastructure operator must respond differently to a sensor anomaly than to a deliberate adversarial manipulation. The paper frames adversarial detection not as an isolated ML security task but as part of a			

Title	Date	Partners	Audience
broader monitoring architecture supporting safe autonomous operation, which aligns closely with infrastructure supervisory control principles. Another strong infrastructure-oriented contribution is the work by Schiendorfer et al., which addresses gas grid management through multi-objective reinforcement learning (MORL) and reports lessons learned from real-world pilot deployments. Their paper emphasizes that infrastructure control is inherently multi-criteria: operators must balance linepack, pressure limits, operational smoothness, and safety constraints, often under changing priorities. Their approach uses the GPI-PD algorithm to maintain a set of policies optimized for different objective weightings, allowing operators to select strategies depending on real-time needs (Schiendorfer et al., 2025). The paper's main practical insight is that deployment is limited not by the availability of RL algorithms but by infrastructure realities: the effort required for labeling, the difficulty of generating negative training examples, the scalability limitations of high-fidelity simulation, and the importance of interpretability and trust for dispatchers. Their lessons highlight a key challenge for critical infrastructure: AI deployment is fundamentally constrained by data availability, operational acceptance, and the ability to integrate AI into existing SCADA and decision-support workflows (Schiendorfer et al., 2025). This paper demonstrates that the true bottleneck is not "learning a policy," but rather building an ecosystem in which a policy can be trusted and adopted. Complementing these technical works, the paper on continuous assessment-driven requirement elicitation (Fedorova et al., 2025) makes an important contribution to governance and engineering processes. It adapts the European Commission's ALTAI framework into a lifecycle method that distinguishes ethical requirements relevant			

Title	Date	Partners	Audience
at different Technology Readiness Levels (TRLs), enabling a staged approach to ethics-by-design. Their railway management use case shows how ALTAI questions can be converted into functional and non-functional requirements, while explicitly classifying which issues are relevant for proof-of-concept and which will become relevant in full-scale deployment. For critical infrastructures, this is particularly important because infrastructure AI systems are rarely deployed in a single step; they evolve from prototypes to operational systems, and ethical risks often shift across this maturity trajectory. Their contribution highlights a challenge that is often underestimated: assessment frameworks designed for ex-post evaluation may foster false confidence when used prematurely. The authors therefore argue that trustworthy AI requires continuous assessment, aligned with regulatory requirements such as the EU AI Act's lifecycle risk management obligations. This paper reinforces the idea that trustworthiness is not solely a technical attribute but a development discipline, linking requirements engineering to compliance and operational readiness. From these contributions, several promising research lines become evident. One major direction is the development of architectural trust patterns for AI components, extending beyond monitoring into compositional assurance. Atif et al.'s call to design for trust rather than correctness suggests a need for reusable architectural templates for critical infrastructure, including runtime monitors, redundancy mechanisms, reject options, and fail-safe strategies that are compatible with certification. This intersects directly with Lowe et al.'s work, where detection and supervision architectures become key. A future research line would therefore focus on unifying monitoring approaches for uncertainty, anomalies, and adversarial threats into a standardized supervisory layer			

Title	Date	Partners	Audience
for AI-enabled infrastructure systems. This could lead to reference architectures that integrate epistemic uncertainty estimation, distribution shift detection, adversarial attribution, and safe fallback policies, with explicit interfaces to human operators. A second research line concerns safe reinforcement learning training methodologies that avoid the paradox highlighted by King et al.: shielding improves safety but may prevent the agent from learning. This suggests a need for adaptive shielding strategies, staged curricula, or dual-mode learning approaches where safety constraints are gradually tightened as policies mature. The checklist-based framework for safe RL suggests the importance of structured design workflows, but further research is needed to operationalize these workflows into toolchains that infrastructure operators can apply. The MORL pilot deployment in gas grids also suggests that RL must be embedded in decision-support paradigms rather than replacing dispatchers. Research could therefore explore interactive RL and MORL frameworks in which policy generation is automated but selection and accountability remain human-centered, aligning with "human-in-command" infrastructure principles. A third research line involves data generation and simulation fidelity. Both gas grid and cruise ship papers show that realistic simulators are too slow or incomplete, while purely data-driven surrogate models have blind spots due to limited state coverage. A promising direction is hybrid simulation architectures combining physics-based models, surrogate neural regressors, and scenario generation engines that produce rare failure states. This is essential for infrastructures because historical data is biased toward safe operation, leaving insufficient negative examples for learning safe responses. Methods such as synthetic stress testing, counterfactual simulation, and			

Title	Date	Partners	Audience
adversarial scenario generation could become core components of infrastructure AI validation pipelines. A fourth research line concerns requirements engineering and governance mechanisms that scale. The ALTAI adaptation paper demonstrates that trustworthiness frameworks must be lifecycle-sensitive. Future research could focus on formalizing mappings between ethical assessment checklists, infrastructure safety constraints, and verification artifacts such as safety cases. This could produce a bridge between EU regulatory compliance and technical design, enabling infrastructure operators to build auditable evidence chains linking requirements, training data decisions, algorithmic constraints, monitoring policies, and operational logs. Such work is crucial because infrastructures must demonstrate not only performance but also accountability, explainability, and controlled risk exposure. The synergies between the papers are particularly strong and point to an emerging integrated vision. The concept of "unknown inputs" and trust-based architecture from Atif et al. aligns naturally with Lowe et al.'s work on distinguishing adversarial from natural anomalies, since both are fundamentally about detecting when the AI system is operating outside its validated regime. Their approaches could be unified: unknown-input handling could be enriched with threat classification mechanisms, enabling infrastructure systems to not only reject uncertain outputs but also infer whether the uncertainty is environmental or malicious. Similarly, the STPA-driven safety shielding approach aligns closely with the safe RL checklist paper: shielding is explicitly identified as a safety layer, and STPA provides a systematic method for deriving the constraints that shielding must enforce. This provides a bridge between hazard analysis and RL engineering, which is essential in			

Title	Date	Partners	Audience
infrastructure where safety constraints must be justified at the system level. There is also a strong synergy between the gas grid MORL deployment and the ALTAI-driven requirement elicitation paper. The gas grid work highlights the practical difficulty of labeling and operator trust, while the ALTAI-based method provides a structured mechanism to elicit requirements for transparency, accountability, and fairness early in the design. Together, they suggest a holistic infrastructure workflow: ethical and trustworthiness requirements are elicited continuously across TRL stages, then translated into system constraints, monitoring requirements, logging policies, and decision support interfaces that operators can understand and accept. In this sense, MORL multi-policy approaches can serve as a practical implementation of “human agency and oversight,” since operators can choose among policies rather than blindly executing a single learned policy. Finally, the workshop shows that trustworthiness in critical infrastructures must be approached as a multi-layered ecosystem problem, where algorithmic advances alone are insufficient. Robustness to uncertainty, safety enforcement, monitoring and anomaly detection, multi-objective decision-making, and lifecycle requirement elicitation all form interdependent layers. The synergy is that each paper contributes a piece of a unified stack: requirements and governance (ALTAI-based elicitation), architectural trust patterns (Beyond Correctness), monitoring and adversarial differentiation (Mahalanobis-based detection), RL design guidelines (checklists and definitions), safety enforcement mechanisms (STPA-based shields), and operational deployment experience (gas grid MORL pilots). Taken together, they suggest a mature research agenda: critical infrastructures require AI systems that are not only accurate, but also			

Title	Date	Partners	Audience
continuously assessed, architected for uncertainty, safeguarded by explicit constraints, supervised against anomalies and attacks, and designed to preserve human authority. Atif, M., Zoppi, T., & Bondavalli, A. (2025). Beyond correctness: Architecting trustworthy software for autonomous systems in the age of AI. ECML PKDD 2025, Porto, Portugal. Fedorova, A., Stefani, W., Heitz, C., Chavarriaga, R. (2025). Continuous assessment-driven requirement elicitation for trustworthy AI systems. (2025). ECML PKDD 2025, Porto, Portugal. Yamagata, T., Santos-Rodriguez, R. (2025). Guidelines for safe and robust reinforcement learning: From definitions to design. ECML PKDD 2025, Porto, Portugal. King, A., Shahinas, E., & Atmojo, U. (2025). Constructing safety shields for reinforcement learning agents with system-theoretic process analysis. ECML PKDD 2025, Porto, Portugal. Lowe, R., Ulan, M., Bui, T. H., Giordana, G., Giani, T., & Rossi, F. (2025). Differentiating adversarial attacks from natural sensory anomalies in object detection. ECML PKDD 2025, Porto, Portugal. Schiendorfer, A., Schmidt, T., Outafraout, K., Baschin, M., Hei, Y., Streubel, T., Kätzel, P., Michailov, L., Felix, R., & Harpeng, L. (2025). Multi-objective reinforcement learning for safety-critical gas grid management: Lessons from real-world pilot deployment. WECML PKDD 2025, Porto, Portugal. **ANNEX 4 – IHIET 2025 workshop on Human-**			

Title	Date	Partners	Audience
AI Collaboration in Critical Tasks			
Workshop “Machine Learning for Sustainable Power Systems Workshop” @ECMLPKDD 2025	September 15, 2025	Fraunhofer, RTE, UKassel	35 participants
Human-AI Collaboration in Critical Tasks (Special Session @IHIET International Conference 2025) – see a summary of conclusions of this workshop in ANNEX 3 – ECML PKDD 2025 workshop on critical infrastructures The set of papers presented at the workshop shows that trustworthy AI for critical infrastructures is heading away from isolated algorithmic performance claims and towards system-level assurance, where safety, robustness, accountability, and operational feasibility must be engineered together. Across the contributions, a recurring theme emerges: critical infrastructures cannot afford to treat AI as a “smart component” inserted into an otherwise deterministic control chain. Instead, AI must be embedded as a managed and supervised element of a socio-technical system, continuously evaluated, constrained, and aligned with safety envelopes and human operational priorities. A central contribution of the workshop papers is the explicit recognition that traditional correctness is insufficient for autonomous or AI-driven decision-making in critical infrastructures. Atif et al. argue that complex ML-based functions inherently face “unknown inputs” that cannot be exhaustively tested, thus undermining classical verification approaches and certification logic. Their position reframes the problem: infrastructures should not demand perfect correctness, but rather trustworthiness, supported through architectural patterns such as monitoring, redundancy, rejection mechanisms, and diversity (Atif et al., 2025). This is particularly relevant for infrastructures such as transportation networks, energy systems, or	August 26, 2025	FHNW	n/a

Title	Date	Partners	Audience
automated surveillance, where operational environments are open-ended, and the cost of failure is high. Their argument introduces a key challenge: infrastructures need engineering methods to make systems dependable even when components behave unpredictably, rather than attempting to eliminate unpredictability entirely (Atif et al., 2025). This is a major shift in how safety cases for AI-enabled infrastructure control can be constructed. A second important contribution lies in the reinforcement learning (RL) domain, where several papers highlight both the promise and the fragility of RL for infrastructure control. King et al. explore a cruise ship HVAC optimization scenario in which RL is used to improve energy efficiency under uncertain environmental and operational conditions. Their work introduces an important link between classical hazard analysis and RL safety: they use System-Theoretic Process Analysis (STPA) to identify hazards and translate them into constraints and controller requirements, then implement these constraints and requirements as a safety shield that blocks unsafe actions (King et al., 2025). The study reveals a practical challenge: shielding improves safety during deployment, but blocking unsafe actions during training can hinder the RL agent's ability to learn a safe policy. This reflects a critical infrastructure dilemma: safety enforcement mechanisms may distort learning dynamics, creating a tension between exploration and constraint satisfaction that is rarely acknowledged in theoretical RL research. Their paper thus contributes an applied insight: safety must be treated not only as a deployment property, but also as a training-time design variable, particularly when infrastructures demand predictable behavior from the first day of operation (King et al., 2025). This challenge is further reinforced by the broader perspective provided by (Yamagata			

Title	Date	Partners	Audience
et al., 2025), which addresses the fragmentation of terminology and methodology in safe RL. The authors propose consensus definitions and a structured checklist spanning ethical requirements, reward specification, robustness against uncertainty, explicit constraints, the use of simulators and expert knowledge, and finally, safety layers such as shielding and human intervention. For critical infrastructures, this paper’s key contribution is not a new algorithm but rather a structured engineering viewpoint: safe RL cannot be reduced to constrained optimization alone. It must incorporate traceability, explainability, reward alignment, and mechanisms for human override. The paper also highlights a major infrastructure challenge: the lack of standardized guidelines makes it difficult for operators and regulators to judge whether an RL-based controller is deployable. In safety-critical domains, such ambiguity is itself a risk, because it blocks reproducibility and certification readiness. The workshop also provides evidence that AI safety challenges in infrastructures are increasingly tied to security and adversarial resilience, not only physical failures. Lowe et al. contribute a strong example in the space domain, focusing on satellite object detection. Their work proposes a statistical method based on Mahalanobis distance to distinguish between adversarial perturbations and natural sensory anomalies. Their findings show that adversarial attacks can degrade detection performance at perturbation levels where natural noise does not trigger detection, and that common anomaly detection methods such as variational autoencoders may fail to detect adversarial manipulations (Lowe et al., 2025). This contribution is highly relevant to critical infrastructure because it highlights the necessity of threat differentiation: an infrastructure operator must respond differently to a sensor anomaly than to a			

Title	Date	Partners	Audience
deliberate adversarial manipulation. The paper frames adversarial detection not as an isolated ML security task but as part of a broader monitoring architecture supporting safe autonomous operation, which aligns closely with infrastructure supervisory control principles. Another strong infrastructure-oriented contribution is the work by Schiendorfer et al., which addresses gas grid management through multi-objective reinforcement learning (MORL) and reports lessons learned from real-world pilot deployments. Their paper emphasizes that infrastructure control is inherently multi-criteria: operators must balance linepack, pressure limits, operational smoothness, and safety constraints, often under changing priorities. Their approach uses the GPI-PD algorithm to maintain a set of policies optimized for different objective weightings, allowing operators to select strategies depending on real-time needs (Schiendorfer et al., 2025). The paper's main practical insight is that deployment is limited not by the availability of RL algorithms but by infrastructure realities: the effort required for labeling, the difficulty of generating negative training examples, the scalability limitations of high-fidelity simulation, and the importance of interpretability and trust for dispatchers. Their lessons highlight a key challenge for critical infrastructure: AI deployment is fundamentally constrained by data availability, operational acceptance, and the ability to integrate AI into existing SCADA and decision-support workflows (Schiendorfer et al., 2025). This paper demonstrates that the true bottleneck is not "learning a policy," but rather building an ecosystem in which a policy can be trusted and adopted. Complementing these technical works, the paper on continuous assessment-driven requirement elicitation (Fedorova et al., 2025) makes an important contribution to governance and engineering processes. It			

Title	Date	Partners	Audience
adapts the European Commission’s ALTAI framework into a lifecycle method that distinguishes ethical requirements relevant at different Technology Readiness Levels (TRLs), enabling a staged approach to ethics-by-design. Their railway management use case shows how ALTAI questions can be converted into functional and non-functional requirements, while explicitly classifying which issues are relevant for proof-of-concept and which will become relevant in full-scale deployment. For critical infrastructures, this is particularly important because infrastructure AI systems are rarely deployed in a single step; they evolve from prototypes to operational systems, and ethical risks often shift across this maturity trajectory. Their contribution highlights a challenge that is often underestimated: assessment frameworks designed for ex-post evaluation may foster false confidence when used prematurely. The authors therefore argue that trustworthy AI requires continuous assessment, aligned with regulatory requirements such as the EU AI Act’s lifecycle risk management obligations. This paper reinforces the idea that trustworthiness is not solely a technical attribute but a development discipline, linking requirements engineering to compliance and operational readiness. From these contributions, several promising research lines become evident. One major direction is the development of architectural trust patterns for AI components, extending beyond monitoring into compositional assurance. Atif et al.’s call to design for trust rather than correctness suggests a need for reusable architectural templates for critical infrastructure, including runtime monitors, redundancy mechanisms, reject options, and fail-safe strategies that are compatible with certification. This intersects directly with Lowe et al.’s work, where detection and supervision architectures become key. A future research line would therefore focus on			

Title	Date	Partners	Audience
unifying monitoring approaches for uncertainty, anomalies, and adversarial threats into a standardized supervisory layer for AI-enabled infrastructure systems. This could lead to reference architectures that integrate epistemic uncertainty estimation, distribution shift detection, adversarial attribution, and safe fallback policies, with explicit interfaces to human operators. A second research line concerns safe reinforcement learning training methodologies that avoid the paradox highlighted by King et al.: shielding improves safety but may prevent the agent from learning. This suggests a need for adaptive shielding strategies, staged curricula, or dual-mode learning approaches where safety constraints are gradually tightened as policies mature. The checklist-based framework for safe RL suggests the importance of structured design workflows, but further research is needed to operationalize these workflows into toolchains that infrastructure operators can apply. The MORL pilot deployment in gas grids also suggests that RL must be embedded in decision-support paradigms rather than replacing dispatchers. Research could therefore explore interactive RL and MORL frameworks in which policy generation is automated but selection and accountability remain human-centered, aligning with “human-in-command” infrastructure principles. A third research line involves data generation and simulation fidelity. Both gas grid and cruise ship papers show that realistic simulators are too slow or incomplete, while purely data-driven surrogate models have blind spots due to limited state coverage. A promising direction is hybrid simulation architectures combining physics-based models, surrogate neural regressors, and scenario generation engines that produce rare failure states. This is essential for infrastructures because historical data is biased toward safe operation, leaving			

Title	Date	Partners	Audience
insufficient negative examples for learning safe responses. Methods such as synthetic stress testing, counterfactual simulation, and adversarial scenario generation could become core components of infrastructure AI validation pipelines. A fourth research line concerns requirements engineering and governance mechanisms that scale. The ALTAI adaptation paper demonstrates that trustworthiness frameworks must be lifecycle-sensitive. Future research could focus on formalizing mappings between ethical assessment checklists, infrastructure safety constraints, and verification artifacts such as safety cases. This could produce a bridge between EU regulatory compliance and technical design, enabling infrastructure operators to build auditable evidence chains linking requirements, training data decisions, algorithmic constraints, monitoring policies, and operational logs. Such work is crucial because infrastructures must demonstrate not only performance but also accountability, explainability, and controlled risk exposure. The synergies between the papers are particularly strong and point to an emerging integrated vision. The concept of “unknown inputs” and trust-based architecture from Atif et al. aligns naturally with Lowe et al.’s work on distinguishing adversarial from natural anomalies, since both are fundamentally about detecting when the AI system is operating outside its validated regime. Their approaches could be unified: unknown-input handling could be enriched with threat classification mechanisms, enabling infrastructure systems to not only reject uncertain outputs but also infer whether the uncertainty is environmental or malicious. Similarly, the STPA-driven safety shielding approach aligns closely with the safe RL checklist paper: shielding is explicitly identified as a safety layer, and STPA provides a systematic method for deriving the			

Title	Date	Partners	Audience
constraints that shielding must enforce. This provides a bridge between hazard analysis and RL engineering, which is essential in infrastructure where safety constraints must be justified at the system level. There is also a strong synergy between the gas grid MORL deployment and the ALTAI-driven requirement elicitation paper. The gas grid work highlights the practical difficulty of labeling and operator trust, while the ALTAI-based method provides a structured mechanism to elicit requirements for transparency, accountability, and fairness early in the design. Together, they suggest a holistic infrastructure workflow: ethical and trustworthiness requirements are elicited continuously across TRL stages, then translated into system constraints, monitoring requirements, logging policies, and decision support interfaces that operators can understand and accept. In this sense, MORL multi-policy approaches can serve as a practical implementation of “human agency and oversight,” since operators can choose among policies rather than blindly executing a single learned policy. Finally, the workshop shows that trustworthiness in critical infrastructures must be approached as a multi-layered ecosystem problem, where algorithmic advances alone are insufficient. Robustness to uncertainty, safety enforcement, monitoring and anomaly detection, multi-objective decision-making, and lifecycle requirement elicitation all form interdependent layers. The synergy is that each paper contributes a piece of a unified stack: requirements and governance (ALTAI-based elicitation), architectural trust patterns (Beyond Correctness), monitoring and adversarial differentiation (Mahalanobis-based detection), RL design guidelines (checklists and definitions), safety enforcement mechanisms (STPA-based shields), and operational deployment experience (gas grid MORL pilots). Taken			

Title	Date	Partners	Audience
together, they suggest a mature research agenda: critical infrastructures require AI systems that are not only accurate, but also continuously assessed, architected for uncertainty, safeguarded by explicit constraints, supervised against anomalies and attacks, and designed to preserve human authority. Atif, M., Zoppi, T., & Bondavalli, A. (2025). Beyond correctness: Architecting trustworthy software for autonomous systems in the age of AI. ECML PKDD 2025, Porto, Portugal. Fedorova, A., Stefani, W., Heitz, C., Chavarriaga, R. (2025). Continuous assessment-driven requirement elicitation for trustworthy AI systems. (2025). ECML PKDD 2025, Porto, Portugal. Yamagata, T., Santos-Rodriguez, R. (2025). Guidelines for safe and robust reinforcement learning: From definitions to design. ECML PKDD 2025, Porto, Portugal. King, A., Shahinas, E., & Atmojo, U. (2025). Constructing safety shields for reinforcement learning agents with system-theoretic process analysis. ECML PKDD 2025, Porto, Portugal. Lowe, R., Ulan, M., Bui, T. H., Giordana, G., Giani, T., & Rossi, F. (2025). Differentiating adversarial attacks from natural sensory anomalies in object detection. ECML PKDD 2025, Porto, Portugal. Schiendorfer, A., Schmidt, T., Outafraout, K., Baschin, M., Hei, Y., Streubel, T., Kätzel, P., Michailov, L., Felix, R., & Harpeng, L. (2025). Multi-objective reinforcement learning for safety-critical gas grid management: Lessons from real-world pilot deployment. WECML PKDD 2025, Porto, Portugal.			

Title	Date	Partners	Audience
ANNEX 4 – IHIET 2025 workshop on Human-AI Collaboration in Critical Tasks			
Hackathon Hack4rail	September 23-25, 2025	SBB, DB, FLATLAND	100 participants
Flatland Workshop and Symposium 2025	November 17-19, 2025	FLATLAND, SBB, FHNW, ENLITEAI	32 participants
Workshop "AI technologies for monitoring workplaces and occupational health"	March 11, 2026	INESC TEC	20 participants

TABLE 8 - EVENTS ORGANISED BY THE PROJECT'S CONSORTIUM

In addition to these events, one can also count the organisation of an internal event that aimed only the consortium partners, which was the following:

- AI4REALNET Third Consortium Meeting | June 3-4, 2025 [40 participants]

Alongside the events organised by the consortium, 26 events were attended by some of the project's partners. Figure 9 - Percentage of events by type attended by the project's Consortium shows the percentage of events by type attended by the project's consortium, as Table 9 - Events attended by the project's consortium presents a detailed list of those events.

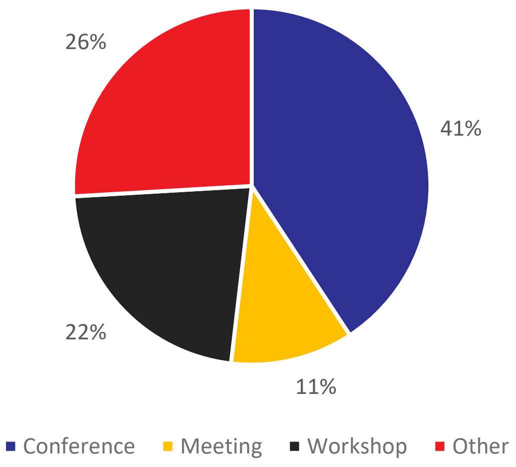


FIGURE 9 - PERCENTAGE OF EVENTS BY TYPE ATTENDED BY THE PROJECT'S CONSORTIUM

Title	Date	Participation	Partners	Audience
European Convergence Summit (ECS) 2025	April 9, 2025	Project presentation (2 slides)	INESC TEC	n/a
SwissNLP | Expert Group: Responsible and Secure NLP	March 27, 2025	NLP Expert Group Meeting	ZHAW	n/a
ZHAW Sustainability Day 2025	April 8, 2025	Event	ZHAW	n/a
Psychological Aspects of Human-Machine Interaction @The Applied Machine Learning Days (AMLD) EPFL 202	February 11-14, 2025	Presentation	FHNW	n/a
Reinforcement learning for real-world network infrastructure @Workshop on Modern Applications of Control Theory and Reinforcement Learning	May 20, 2025	Presentation	UvA	75 participants
Deep Reinforcement Learning for Combinatorial Optimization @4th workshop on Mathematics and AI	June 12, 2025	Presentation	UvA	50 participants
"Human-AI Co-creation: How can we bring the best of human and machine together?" @Marie Skłodowska-Curie Actions lunchtime conversation	May 23, 2025	Presentation	INESC TEC	n/a
LF Energy Summit Europe 2025	September 10-11, 2025	Project presentation	INESC TEC	40 participants
"Trust is Earned, Not Given - Assessing Trustworthiness of Health-Oriented Systems" @7th Digital Health Lab Day	September 3, 2025	Workshop	ZHAW	n/a
International Human Factors Rail Conference 2025	September 18-19, 2025	Poster presentation	SBB, FLATLAND, DB, FHNW	n/a
Eastern European Machine Learning Summer School	July 21-26, 2025	Presentation	UvA	300 participants
"Human Interaction and Emerging Technologies" @Workshop: Human-AI Collaboration in Critical Tasks in IHIET 2025	August 25-27, 2025	Presentation	UvA, TENNET	25 participants
RLEAP Symposium	September 8-10, 2025	Presentation	UvA	30 participants
World Summit on Counterterrorism	September 14-15, 2025	Workshop	ZHAW	25 participants

Title	Date	Participation	Partners	Audience
AI Governance Lecture Series in SS 2025	July 22, 2025	Presentation	ZHAW	40 participants
“Physics informed Neural networks for Power Flow simulations” @ECML PKDD 2025 ML4SPS workshop	September 15, 2025	Presentation	IRTSX	30 participants
“Real-world congestion management of power grids with Artificial Intelligence” @ECML PKDD 2025 AI-SCI workshop	September 15, 2025	Presentation	TENNET	20 participants
“Wallenberg AI, Autonomous Systems, and Software Program”	October 9, 2025	Talk	UvA	40 participants
RTE Colloquium	October 16, 2025	Presentation	FHNW, RTE	20 participants
Web Summit 2025	November 11, 2025	Presentation	INESC TEC	n/a
Transportforum 2026	January 14-15, 2026	Presentation	LiU	1.300 visitors
ATRASS seminar	January 21, 2026	Presentation	IRTSX	35 participants
Swiss Nuclear Forum	November 20, 2025	Presentation	FHNW	n/a
“Primitives for Human-AI Interaction” @Flatland Workshop and Symposium 2025	November 19, 2025	Presentation	FHNW	n/a
“Artificial intelligence and Robotics within Horizon Europe and beyond”	June 19, 2025	Presentation	INESC TEC, IRTSX, ZHAW, FHNW	n/a
AI, Data and Robotics Forum (ADRF25) - AI, Your Way: Trusted, Scalable, and Ready for to Deliver	September 23, 2025	Workshop	INESC TEC	n/a

TABLE 9 - EVENTS ATTENDED BY THE PROJECT’S CONSORTIUM

The organisation of and participation in events is a highly effective means of showcasing the project’s activities to a wide range of audiences. For this reason, such actions are planned to take place regularly throughout the project’s duration, contributing directly to the achievement of its objectives.

2.1.5 ADVISORY BOARD

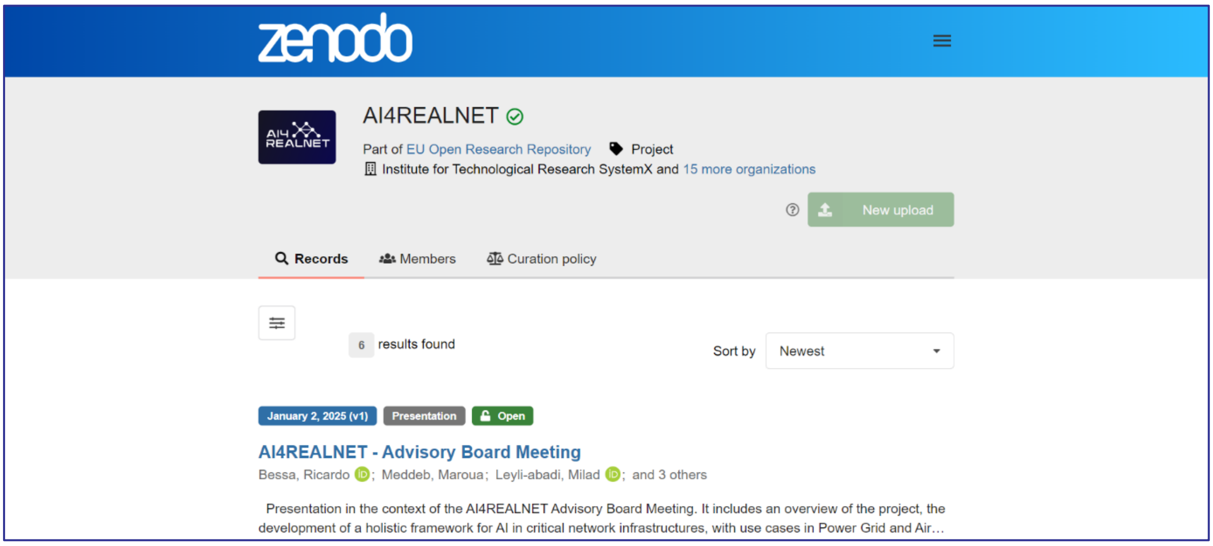
The Advisory Board (AB) consists of 20 confirmed members representing diverse sectors and geographical regions, as detailed in deliverable D5.3. Following the recommendation from the first periodic review, a smaller subgroup of advisers from the AB volunteered to provide more frequent guidance on the project's developments. This dedicated subgroup includes representatives from Intel, DIGI Mind Sphere, TU Dresden, ÖBB-Infrastruktur AG and Artelys as presented in deliverable D5.3.

The Advisory Board continues to play a crucial role in ensuring that the solutions, approaches, and methods developed by AI4REALNET maintain sustained relevance and incorporate current developments at the European level, thereby guaranteeing that project results will have broad applicability and impact across critical infrastructure sectors.

During the reporting period (M19-30), one Advisory Board meeting was held on July 2, 2025 (M22) (see ANNEX 5 – **Advisory Board meeting** for a summary of the meeting outcomes), building on the initial December 2024 (M15) meeting reported in D5.3. The feedback and recommendations from both Advisory Board meetings have been systematically integrated into the project's ongoing work. The insights provided by the AB members have been particularly valuable in refining the evaluation protocols (WP4), strengthening the exploitation strategy (D5.4), and ensuring that the developed solutions address real-world operational needs across the three critical infrastructure domains.

2.1.6 OPEN SCIENCE

The AI4REALNET Zenodo community serves as the central repository for all project research outputs, ensuring their long-term preservation, referral through persistent Digital Object Identifiers (DOIs), and full compliance with the FAIR (Findable, Accessible, Interoperable, Reusable) data management principles. The repository is instrumental in supporting the project's open science strategy, facilitating transparent dissemination and accessibility of project results. The AI4REALNET Zenodo page is available here.



zenodo

AI4REALNET

Part of EU Open Research Repository   Project
Institute for Technological Research SystemX and 15 more organizations

New upload

Records   Members   Curation policy

6 results found

Sort by Newest

January 2, 2025 (v1)   Presentation   Open

AI4REALNET - Advisory Board Meeting

Bessa, Ricardo; Meddeb, Maroua; Leyli-abadi, Milad; and 3 others

Presentation in the context of the AI4REALNET Advisory Board Meeting. It includes an overview of the project, the development of a holistic framework for AI in critical network infrastructures, with use cases in Power Grid and Air…

FIGURE 10 - AI4REALNET ZENODO PAGE

During the first 30 months of implementation, the Zenodo community has demonstrated significant engagement and visibility across academic and professional audiences (see Table 10 - AI4REALNET Zenodo Assets views and download metrics). A total of 18 individual items has been uploaded to the repository, collectively achieving 1.220 unique views and 1.108 unique downloads. The distribution of downloads indicates effective dissemination across asset types. Specifically, publications account for 432 downloads (an average of 43 per paper), followed by presentations with 284 downloads (an average of 71 per item). The datasets have attracted 170 unique downloads, evidencing their use by the wider research community to support further development, testing, and validation activities.

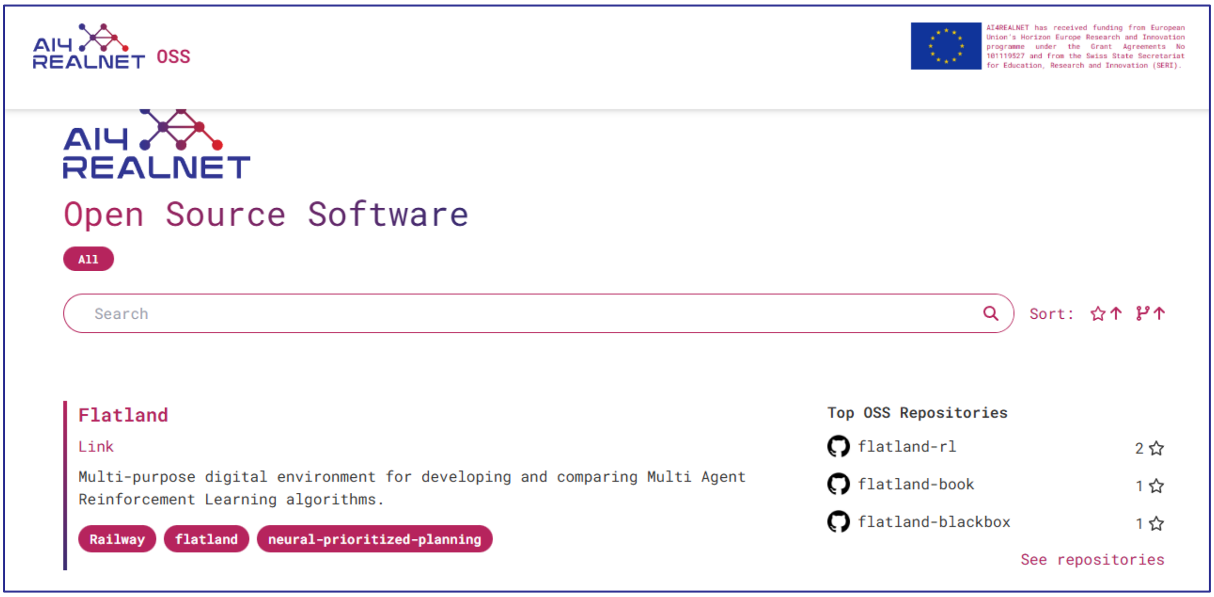
Overall, each deposited item has achieved an average of 61 unique views and 56 unique downloads, demonstrating consistent reach within the artificial intelligence and critical infrastructure research domains. The download-to-view ratio of 91% across all asset types further confirms the high relevance and perceived quality of the outputs made available. These indicators collectively validate the project's data management and open science practices, contributing to the dissemination objectives outlined in the AI4REALNET Data Management Plan and reinforcing the project's broader impact beyond the consortium.

Asset type	Uploads	Unique views	Total views	Unique downloads	Total downloads
Datasets	2	403	431	170	177
Publications	10	223	241	432	481

Asset type	Uploads	Unique views	Total views	Unique downloads	Total downloads
Presentations	4	167	179	284	316
Others	3	365	405	170	196
TOTAL	**19**	**1158**	**1256**	**1056**	**1170**

TABLE 10 - AI4REALNET ZENODO ASSETS VIEWS AND DOWNLOAD METRICS

All project’s key exploitable results are published under an open-source licence and made available through a GitHub repository and page (see Figure 11, https://ai4realnet.github.io/), ensuring concise progress on the AI4REALNET concept, particularly after the end of the project.



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

Open Source Software

Flatland

Link

Multi-purpose digital environment for developing and comparing Multi Agent Reinforcement Learning algorithms.

Top OSS Repositories

flatland-rl	2 ☆
flatland-book	1 ☆
flatland-blackbox	1 ☆

FIGURE 11 - AI4REALNET GITHUB PAGE

All the software on GitHub already has a total of 217 views and 161 contributors, meaning that people are viewing, experimenting, and interacting with the community, giving back information important to the refinement of the results. Note that 66% of those contributors are external to the managing teams (that include people from the AI4REALNET project and INESC TEC). Also, the project's uploads have been cloned 213 times, an indicator that matches the number of views, meaning that the page is engaged and attracting the adequate type of audience.

As the project progresses, the number of scientific publications continues to grow, summing now a total of 29 publications issued and three remain under review. Considering specially the reporting period in question, 13 publications have been produced and disseminated. Table 11 - List of the project's scientific publications published since the beginning of the project presents a detailed list of all the work published within the scope of the AI4REALNET since the beginning of the project, and all available currently in open access.

Title	Date	Venue	Type	Partner(s)
Structured Power Grid Simulation Dataset for Machine Learning: Failure and Survival Events in Grid2Op's L2RPN WCCI 2022 Environment	October 17, 2024	ECML 2024 Workshop	Dataset	Fraunhofer UKASSEL
A Pioneering roadmap for ML-driven algorithmic advancements in electrical networks	October 24, 2024	IEEE ISGT Europe 2024	Conference paper	INESC TEC TUD RTE
Fault Detection for Agents in Power Grid Topology Optimization: A Comprehensive Analysis	August 1, 2024	ECML 2024 Workshop	Conference paper	Fraunhofer UKASSEL
Imitation Learning for Intra-Day Power Grid Operation through Topology Actions	August 18, 2024	ECML 2024 Workshop	Conference paper	TENNET
State and Action Factorization in Power Grids	September 3, 2024	ECML 2024 Workshop	Conference paper	POLIMI
Sub-optimal Experts mitigate Ambiguity in Inverse Reinforcement Learning	December 10, 2024	NeurIPS 2024	Conference paper	POLIMI
How does Inverse RL Scale to Large State Spaces? A Provably Efficient Approach	December 10, 2024	NeurIPS 2024	Conference paper	POLIMI
Last-Iterate Global Convergence of Policy Gradients for Constrained Reinforcement Learning	December 10, 2024	NeurIPS 2024	Conference paper	POLIMI
Graph Reinforcement Learning for Power Grids: A Comprehensive Survey	August 1, 2024	Energy & AI	Journal Paper	Fraunhofer, UKASSEL
Multi-Objective Reinforcement Learning for Power Grid Topology Control	January 27, 2025	IEEE Power Tech 2025 *[29 June-3 July]*	Conference paper	TENNET
Learning Topology Actions for Power Grid Control: A Graph-Based	March 19, 2025	ECML PKDD 2025 *[September 15-19]*	Conference paper	FRAUNHOFER, UKASSEL, TENNET

Title	Date	Venue	Type	Partner(s)
Soft-Label Imitation Learning Approach				
On the Definition of Robustness and Resilience of AI Agents for Real-time Congestion Management	February 14, 2025	IEEE Power Tech 2025 *[29 June-3 July]*	Conference paper	INESC TEC
Generation of Power Network Operating Scenarios for an AI-friendly Digital Environment	February 14, 2025	IEEE Power Tech 2025 *[29 June-3 July]*	Conference paper	INESC TEC, RTE
The Supportive AI Framework: From recommending to supporting	February 6, 2025	HCI International 2025 *[22-27 June]*	Conference paper	TENNET
User experience evaluation of an AI-based decision-support tool for power grid congestion management	May 6, 2025	IHIET 2025 Conference *[25-27 August]*	Conference paper	TENNET, LiU
Centrally Coordinated Multi-Agent Reinforcement Learning for Power Grid Topology Control	February 12, 2025	ACM e-Energy 2025 *[17-20 June]*	Conference paper	TENNET
Investigating Human-AI collaboration in Railway Traffic Management Using Flatland	September 18, 2025	International Human Factors Rail Conference 2025	Poster	SBB, FHNW, FLATLAND, DB
Human-AI interaction in safety-critical network infrastructures	September 19, 2025	iScience	Journal paper	ALL
A Conceptual Framework for AI-based Decision Systems in Critical Infrastructures	October 1, 2025	IEEE SMC 2025 *[5-8 October]*	Conference paper	INESC TEC, IRTSX, TENNET, SBB, TUD, ENLITEAI, ZHAW, FHNW, RTE, FLATLAND, POLIMI, FRAUNHOFER
Study Design and Demystification of Physics Informed Neural Networks for Power Flow Simulation	September 15, 2025	ECML PKDD 2025 *[15-19 September]*	Conference paper	IRTSX, RTE
Power Grid Control with Graph-Based Distributed Reinforcement Learning	September 2, 2025	ECML PKDD 2025 *[15-19 September]*	Conference paper	POLIMI

Title	Date	Venue	Type	Partner(s)
Learned Representations Enhance Multi Agent Path Planning	June 14, 2025	ICML 2025 *[13-19 July]*	Poster	UvA
Continuous Assessment Driven Requirements Elicitation For Trustworthy AI Systems	September 15, 2025	ECML PKDD 2025 *[15-19 September]*	Conference paper	ZHAW
Applying Job Design Criteria for Effective Human-AI Collaboration	August 29, 2025	IHIET 2025 Conference *[25-27 August]*	Conference paper	FHNW
AI-assisted control in network operations. Human-AI teaming in critical infrastructures – a conceptual model	April 30, 2024	HFES Europe Chapter Conference *[9-11 April]*	Poster	FHNW
Evolving power system operator rules for real-time congestion management	January 6, 2026	Energy and AI	Journal paper	INESC TEC
Current and Future Applications of Artificial Intelligence in Power Systems: A Critical Appraisal	February 6, 2026	Journal of Modern Power Systems and Clean Energy	Journal Paper	INESC TEC
"Toward a Holistic Framework for Human-AI Collaboration in Safety-Critical Systems"	March 7, 2026	"Artificial Intelligence, Data and Robotics"	Book chapter	INESC TEC, IRTSX, SBB, TU Delft, Enlite AI, FHNW, NAV, TenneT, Fraunhofer, RTE, DB, LiU, FLATLAND, POLIMI, UvA
"Human-AI Interaction and Visualization Perspectives on ADR"	March 7, 2026	"Artificial Intelligence, Data and Robotics"	Book chapter	LiU

TABLE 11 - LIST OF THE PROJECT’S SCIENTIFIC PUBLICATIONS PUBLISHED SINCE THE BEGINNING OF THE PROJECT

However, we found it relevant to pinpoint some of the work submitted until now, but the process of reviewing is still ongoing. Therefore, Table 12 - List of the project’s scientific publications submitted (under review) presents the list of the work submitted until the date of this deliverable (i.e., under review), but no information on applicable publications has yet been available.

Title	Date of submission	Venue	Type	Partner(s)
Generalizable Graph Neural Networks for Robust Power Grid Topology Control	January 3, 2025	Applied Energy	Journal Paper	TENNET

TABLE 12 - LIST OF THE PROJECT'S SCIENTIFIC PUBLICATIONS SUBMITTED (UNDER REVIEW)

2.1.7 COOPERATION WITH STAKEHOLDERS

Cooperation and synergies with other initiatives are key activities throughout the project timeline, considering the expected results and their dissemination during and after the project. The following subsections summarise the main cooperation initiatives during M19-30.

2.1.7.1 AI-ON-DEMAND PLATFORM

The AI-on-Demand (AIoD) platform unites Europe's diverse AI communities and organisations interested in contributing to or benefiting from AI capabilities, providing a shared space to access resources, tools, and use cases. In this context, AI4REALNET's participation is particularly relevant, as it focuses on AI-based solutions for critical infrastructures such as electricity, railway, and air traffic networks, contributing with open-source environments that support long-term goals in energy transition and digitalisation.

During this reporting period, the consortium continued its efforts to integrate project assets into the AIoD platform. The software assets presented in deliverables D1.3 - "Digital environment – Version 2" and D2.2 - "AI fundamental blocks – beta release", and available in the project GitHub, are progressively being integrated into the AIoD platform. This includes the enhanced versions of the digital environments (Grid2Op, Bluesky, and Flatland) with the project's use cases and data generation assets, replicating real-world operating scenarios involving human operators to apply innovative AI-based methods.

The integration process has experienced some delays due to the ongoing migration of the AIoD platform to a new infrastructure, a factor outside the control of the AI4REALNET consortium. Nevertheless, the project continues to maintain close collaboration with the AIoD team to ensure seamless integration once the new platform infrastructure is fully operational.

2.1.7.2 ADRA

The Adra-e project aims to create the conditions for an inclusive and sustainable European ecosystem in AI, Data and Robotics that strengthens research, innovation and the adoption of trustworthy technologies and applications. Therefore, it is regarded as a key partner for the consortium, which cooperated closely with Adra-e and, during months M19-30 of the project, carrying out multiple joint activities with this key stakeholder.

Participation in ADRF 2025 Event

AI4REALNET participated in the AI, Data and Robotics Forum (ADRF 2025) held in October 2025, with the theme "AI, Your Way: Trusted, Scalable, and Ready to Deliver". The project engaged in collaborative activities with the cluster projects to share insights and best practices on human-centric AI development.

Contribution to ADRA Strategic Research, Innovation, and Deployment Agenda (SRIDA)

The project coordinator, Ricardo Bessa (INESC TEC), actively contributed to the ADRA SRIDA through participation in the Energy Topic Group, which he chairs. This involvement enabled the project to provide strategic input on research challenges and innovation priorities related to AI deployment in critical energy infrastructures. The contributions to SRIDA from AI4REALNET are in **ANNEX 6 –**

Contributions to ADRA SRIDA.

Publication of Book Chapters

Two book chapters submitted to the Adra-e open-access book "Artificial Intelligence, Data and Robotics: Foundations, Transformations, and Future Directions" were published in April 2026:

- First chapter title: "Toward a Holistic Framework for Human-AI Collaboration in Safety-Critical Systems"
- Second chapter title: "Human-AI Interaction and Visualization Perspectives on ADR"

Continued Engagement with ADR Awareness Centre

The Adra-e platform, called ADR Awareness Centre, works as an open repository of ADR educational resources and materials. During this period, the AI4REALNET project continued to publish relevant resources on this platform, ensuring that project outputs remain accessible to the broader AI, Data, and Robotics community.

2.1.7.3 LINUX FOUNDATION

The project established strategic cooperation with the Linux Foundation (LF) Europe to foster open, sustainable, and industry-relevant exploitation of its AI-driven methods, tools, and digital environments. This cooperation aims to ensure long-term impact beyond the project lifetime by embedding selected project outcomes within well-established open-source ecosystems governed by the Linux Foundation, and by structuring exploitable results under coherent dissemination and exploitation identities, including the AINETUS project for power grid and that will be hosted in LF Europe and managed by LF Energy.

A major milestone achieved during the reporting period was the integration of Grid2Op as an open-source project within LF Energy, strengthening collaboration with the broader energy and digital infrastructure communities. This transition represents a strategic step towards ensuring the long-term sustainability and industrial uptake of one of the project’s key digital environments. By hosting Grid2Op in LF Energy, the platform is decoupled from single-organisation ownership and embedded within a trusted, community-driven ecosystem that supports open development, reuse, and contribution from academia, industry, and infrastructure operators.

Overall, cooperation with Linux Foundation Energy reinforces the project’s commitment to open science and open innovation, while ensuring that core software assets - promoted under recognised project identities such as AINETUS - remain accessible, actively maintained, and impactful after the end of EU funding.

2.1.7.4 CLUSTER PROJECTS

In this second reporting period, we continued the connection with the four projects funded under the HORIZON-CL4-2022-HUMAN-02-01 call, focusing on “AI for human empowerment”. The cluster includes the following projects:

- TANGO: Advancing trustworthy AI solutions for manufacturing environments
- THEMIS 5.0: Developing frameworks for human-centric AI implementation in industry
- PEER: Enhancing human-AI collaboration through adaptive systems
- HumAIne: Researching emotional intelligence integration in AI applications

Our collaboration framework consists of regular monthly coordination meetings where we align our communication and dissemination strategies to maximise impact and reach.

Following the work developed in 2024, also in 2025 the group managed to publish a joint work entitled "AI Your Way – Trusted, Scalable & Ready to Deliver" (**ANNEX 7 – Joint new policy brief**), a new policy brief that approaches how Europe's AI ecosystem is shaping a future built on trust, collaboration, and human-centric innovation. This document stemmed from a joint workshop that took place during the ADRA Forum 2025 on September 23, 2025, one of Europe's leading events dedicated to artificial intelligence, data, and robotics, to stress out that the future of AI is not about machines replacing people, but about designing systems that enhance human capabilities, along with an alignment with societal values.

Also, regarding these collective efforts, a series of webinars was launched to raise awareness of the projects involved in this cluster, but mainly to discuss subjects of matter to the common audience, cutting across all the areas worked by these projects, under the motto "less science, more common knowledge". The first webinar, entitled "From Command to Collaboration: Reimaging the human-AI relationship" already took place on January 28, 2026, with an audience of more than 60 participants.

Although this collaboration initially involved the five projects funded under the respective call, outreach efforts have been made to engage additional European projects working on related topics aligned with the expected outcomes. Beyond enabling the exchange of good practices, this approach also enhances the project's visibility at the European level.

2.1.7.5 OTHER PROJECTS

In addition to the work developed within the cluster project's hub, the AI4REALNET signed a memorandum of understanding (MoU) with the ELOQUENCE project in October 2025, a collaboration agreement that aims to foster a connection between these research programs, enhancing coordination, and identifying mutual interests in the field of Artificial Intelligence (AI). This MoU shows a commitment to work together, drawing a preliminary roadmap for a collaboration, providing clarity on expectations and responsibilities for future discussions.

2.1.7.6 DOMAIN-SPECIFIC STAKEHOLDERS

The project places strong emphasis on sustained engagement with domain-specific stakeholders to ensure its work remains impactful and relevant. This section outlines how structured interactions with industry actors, academic partners, and other key stakeholders not only broaden the visibility and dissemination of project outcomes, but also create an effective feedback loop to refine ongoing research and validation activities. By anchoring its developments in real operational needs and concrete sectoral challenges, the project strengthens the practical value, acceptance, and adoption potential of its proposed solutions.

2.1.7.6.1 POWER GRID DOMAIN

During the M19-30 reporting period, AI4REALNET maintained and strengthened its engagement with key power grid stakeholders through participation in major industry and research events.

Linux Foundation Energy Summit 2025

The project presented its Grid2Op integration and broader exploitation strategy at the LF Energy Summit Europe 2025 (September 10-11, 2025), showcasing the transition of this critical digital environment to the Linux Foundation. This presentation reinforced the project's commitment to open, sustainable, and community-driven development of power grid AI tools.

ECML PKDD 2025 Conference

The project made significant contributions to the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2025), presenting a relevant workshop to power grid operations:

- “Machine Learning for Sustainable Power Systems” (September 15, 2025), addressing several key topics within the energy system such as Grid Control and Optimization, Predictive Maintenance, Energy Markets, and Ethical AI and Socioeconomic Impacts, among others.

IEEE Spectrum

AI4REALNET received significant visibility in IEEE Spectrum, a prestigious technology magazine dedicated to engineering and the applied sciences. The publication featured Ricardo Bessa from INESC TEC, the AI4REALNET project coordinator, in an analysis of the April 2025 Iberian blackout that affected Spain and Portugal. The article examined how AI4REALNET's high-speed AI decision-making support tools aim to provide grid operators with enhanced capabilities to prevent cascading events and blackouts when critical failures occur. The article also highlighted how AI4REALNET's research tools can help grid operators and researchers understand why failures occur much faster than today's investigative methods, thereby supporting post-event analysis and operational learning across critical infrastructure networks.

2.1.7.6.2 RAILWAY DOMAIN

AI4REALNET expanded its engagement with railway stakeholders during M19-30 through both organised events and participation in industry forums.

Hack4Rail Hackathon 2025

The project organised the Hack4Rail 2025 hackathon (September 23-25, 2025), which focused on developing open-source prototypes within the Flatland environment to simulate realistic disruption scenarios. This event engaged 100 participants from the broader AI and railway community in collaborative problem-solving and innovation, while promoting the use of the project's digital environment for addressing real-world railway operational challenges. Flatland ran the challenge at this event with ~12 participants. A summary of the outcomes is presented in ANNEX 8 –

Hack4Rail 2025 challenge.

International Human Factors Rail Conference 2025

The project contributed to the International Human Factors Rail Conference 2025 (September 18-19, 2025) with a poster presentation titled "Investigating Human-AI collaboration in Railway Traffic Management Using Flatland" by partners SBB, FLATLAND, DB, and FHNW. This contribution disseminated findings on human factors considerations in the design and deployment of AI-assisted railway operations.

Flatland Workshop and Symposium 2025

The Flatland Workshop and Symposium 2025 (November 17-19, 2025) brought together researchers, developers, and industry professionals to advance the state-of-the-art in railway AI applications. Project partners presented on "Primitives for Human-AI Interaction" (November 19, 2025), sharing insights on designing human-centered interfaces for AI-assisted railway dispatching. The event fostered collaborative problem-solving on key technical challenges such as scalability, graph-based environment representations, and human-machine interaction design.

2.1.7.6.3 AIR TRAFFIC MANAGEMENT DOMAIN

The consultation with Maastricht Upper Area Control Centre (MUAC), initiated in the previous reporting period, continued to provide valuable feedback on the AI tools being developed by AI4REALNET. MUAC operates in an environment with similar operational conditions to NAV Portugal, facing the same challenges of aviation growth. Representatives from MUAC participate in the Advisory Board of AI4REALNET, maintaining an ongoing interest in AI applications for the air traffic management domain use cases.

The ongoing dialogue has confirmed the direction of the research undertaken by AI4REALNET in air traffic management and continues to provide opportunities for feedback on the AI tools developed throughout the project.

2.2 KEY PERFORMANCE INDICATORS SUMMARY

To measure the success of each communication action, we set up Key Performance Indicators (KPIs). These indicators, as defined at the time of the proposal, are crucial to evaluating the course of action. Table 13 - AI4REALNET Communication KPIs presents the KPIs associated with the actions considered within the communication and dissemination plan.

Activity	Schedule	KPI	Total progress	% achieved
Website	M3-M42	> 1500/year unique visitors	6.530 visitors	100%
Social media	M1-M42	> 100 followers (LinkedIn) > 100 followers (X) > 100 views (YouTube)	883 followers (LinkedIn) 87 followers (X) 1.448 views (YouTube)	100% 86% 100%
Brochures	M5-42	≥ 5	3	60%
Project Slide-deck	M5	1	1	100%
Videos	M6-M42	≥ 3	26	100%
Infographics	M5	≥ 8	6	75%
Articles	M1-M42	≥ 5	3	60%
Newsletters	M4-M42	≥ 9	7	67%
Press Relations	M2-M42	≥ 3 PRs ≥ 10 news pieces	3 11	100% 100%
Invited talks	M1-M42	On invite	6	n/a
Booklet	M42	1	0	0
Final event	M42	≥ 100 attendees	0	0

TABLE 13 - AI4REALNET COMMUNICATION KPIS

Table 14 - AI4REALNET Dissemination KPIs provides an overview of the achievements for the remaining KPIs over the past 30 months. In terms of publications in highly ranked international journals, the project has produced three papers accepted in Q1 journals, including a perspective article

in iScience (Cell Press). Several additional manuscripts are currently under review with the editorial boards of other leading international journals.

Under the AI4REALNET project, partners contributed 29 research papers to international conferences and journals, including ECML PKDD 2024, IEEE ISGT Europe 2024, NeurIPS 2024, ECML PKDD 2025, IHIET 2025, ACM e-Energy 2025, HCI International 2025, and IEEE SMC 2025. During the reporting period, 13 conference contributions were added, bringing the project to 95% achievement of the target of 20 contributions.

Thematic workshops are a valuable way to foster knowledge exchange, networking, stakeholder engagement, and dissemination. The project successfully achieved its target of organising five workshops (two in-person + three online):

- ECML PKDD 2025 (September 2025): Two workshops organised:
    - "AI for Safety-Critical Infrastructures" (in-person, ~20 participants)
    - "Machine Learning for Sustainable Power Systems" (in-person, ~30 participants)
- IHIET 2025 (August 2025): Special Session "Human-AI Collaboration in Critical Tasks" (~30 participants)
- Additional webinars during this period:
    - "Knowledge-Assisted AI Applications for Real-World Network Infrastructure" (34 participants)

Additionally, the project organised the Hack4Rail hackathon (100 participants) and a track at AMLD EPFL 2025 entitled "AI and Simulation: Solving Complex Real-World Challenges" (45 participants).

As for other engagement activities, the project had a project pitch in a meeting with policymakers (Federal Ministry for Digital and Transport) at the mFUND Workshop Series about Mobility Innovation, and the project coordinator chaired sessions at ADRF 2024 and ADRF 2025 on critical infrastructure management. Within the Adra Forum 2025, the project coordinator also participated in a session about ApplyAI strategy with the European Commission, entitled "GenAI4EU within the Apply AI Strategy".

Regarding AI open innovation competitions, the project plans to organise at least three competitions (Power Grid in Spring 2026, Railway and Air Traffic Management in Fall 2026). Organising these AI open competitions takes approximately 6 to 9 months, from preparation through to the end of the challenge. The Railway competition planning commenced in M17 with the kick-off and discussion of general concepts. The consortium continued defining the problem, set a detailed timeline, and initiated scenario evaluation and challenge setup, with the competition expected to launch in late 2025. However, this was delayed due to difficulties in alignment with existing conferences.

The collaboration with the cluster of European projects continued during this period, maintaining regular coordination meetings with TANGO, THEMIS 5.0, PEER, and HumAIne projects. The consortium also expanded engagement to include collaboration with the Eloquence project and maintained active participation in ADRA-e activities, contributing to the Strategic Research, Innovation, and Deployment Agenda (SRIDA) and publishing two book chapters in the ADRA-e open-access book.

Regarding targeted meetings with policymakers, the project coordinator participated and chaired the ADRA Topics Group session focused on "Inspection and Maintenance & Energy," discussing research challenges in deploying AI to manage critical infrastructures. This complements the earlier engagement at the mFUND Workshop Series. Additionally, at the ADRA Forum 2025 (September 23-24, 2025, in Stavanger, Norway), Ricardo Bessa served as the invited speaker representing the ADRA thematic group on energy during the GenAI4EU within the Apply AI Strategy session. This prestigious forum engagement linked the GenAI4EU initiative - a European Commission programme designed to mobilize generative AI models to serve the needs of Europe's public sector and industry - with the ApplyAI strategy, which seeks to foster real-world AI applications in strategic sectors. These engagements demonstrate AI4REALNET's active role in shaping European AI policy discussions and bring the total policymaker engagement activities to three (75% of target), exceeding the initial projection.

Overall, the project has demonstrated strong performance across communication and dissemination KPIs, with several targets already exceeded (website traffic, social media engagement, videos, press relations, workshops) and others progressing steadily towards completion by M42. The systematic approach to dissemination and the active engagement with multiple stakeholder groups have contributed to raising awareness of the project’s objectives and outcomes at international, European, and domain-specific levels.

Activity	KPI	Total progress	% achieved
AI open innovation competitions (including tutorials)	≥ 3	0	0
Publication in highly ranked international journals	≥ 8	4	50%
Thematic workshops organisation	≥ 5 (2 in person + 3 online)	5	100%

Contributions to international conferences	≥ 20	22	110%
Cluster of European projects and other initiatives	≥ 8	6	75%
Targeted meetings with policymakers	≥ 4	3	75%

TABLE 14 - AI4REALNET DISSEMINATION KPIS

3. PROJECT COMMUNICATION

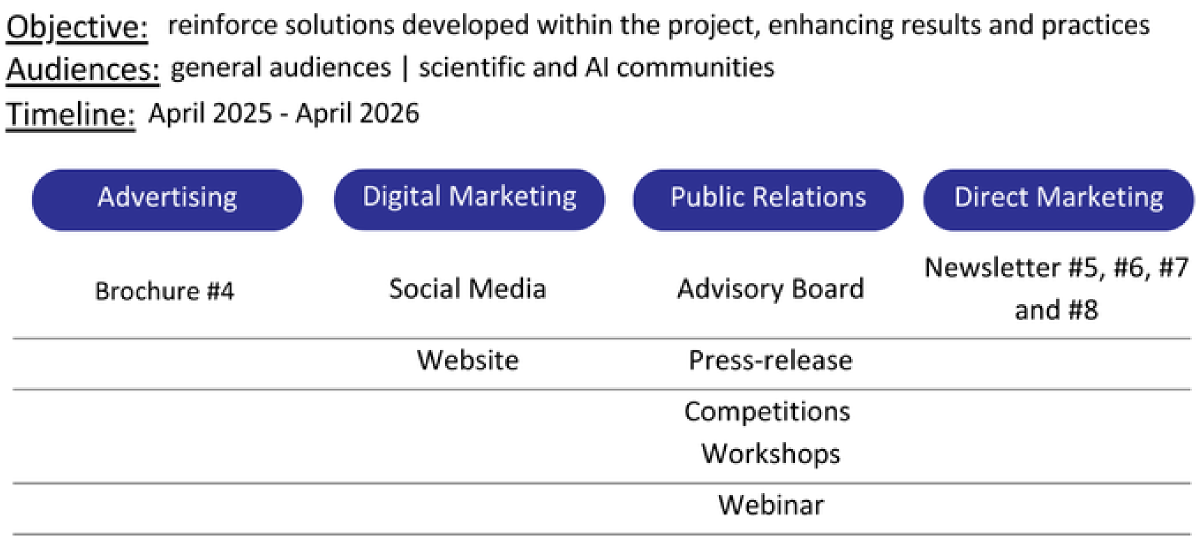
The communication plan outlined in D5.1 was organised into four distinct campaigns to more effectively address the overall objectives of the WP, with each campaign defined by its own goals, target audiences, and communication tools.

The first communication campaign, focused on delivering strategies to create awareness of the project's initiatives, took place between October 2023 (beginning of the project) and April 2024. In turn, the second communication campaign worked to promote the activities developed by the project's partners, deliver relevant input to scientific communities, and engage with AI communities. This happened between May 2024 and March 2025 (M8-18). All the results collected through these two campaigns are available in D5.3 – Communication and dissemination monitoring phase 1.

Accordingly, this section focuses now on presenting the communication outcomes achieved within the campaign implemented between months M19 and M30 of the project.

3.1 THIRD CAMPAIGN RESULTS

The third communication campaign goes from April 2025 until April 2026, covering months M19 to M30. As described on the initial communication and dissemination plan, a summary of this communication campaign – goal, stakeholders, timeline and communication tools – is presented in the next figure:



Objective: reinforce solutions developed within the project, enhancing results and practices
Audiences: general audiences | scientific and AI communities
Timeline: April 2025 - April 2026

Advertising	Digital Marketing	Public Relations	Direct Marketing
Brochure #4	Social Media	Advisory Board	Newsletter #5, #6, #7 and #8
	Website	Press-release	
		Competitions Workshops	
		Webinar	

FIGURE 12 - SUMMARY OF THE THIRD COMMUNICATION CAMPAIGN

Following the successful results achieved with the first and second communication campaigns, the third communication campaign also kept up with the good results. Not all planned actions were executed, as the results are in the Table 15 - Communication results of the third communication campaign. However, the team managed to overcome many of the purposed KPIs. For example, the KPIs established until the end of the project for the website, featuring the page visitors, have already been achieved, although the project is only halfway done, alongside the KPIs already achieved at the time of the previous monitoring report, with the videos.

The brochure edition in this reporting period was #2 and #3, not yet reaching the #4. The delay comes from the previous communication monitoring and will remain until the end of the project, since the team felt the need to re-evaluate and adjust the contents for the brochures, and a revision of the frequency of these publications was implemented. The KPIs will be achieved, but not according to the dates initially planned. The same regarding the press release planned for this campaign. The KPI has already been achieved in this indicator; however, there will be more valuable information to disseminate through the media further along the project, which explains the decision to postpone this release.

Another relevant outcome refers to the social media efforts, especially on what concerns the profile on LinkedIn, where the numbers have reached eight times what was proposed for the entire project. The results on the X platform have been slowly increasing, even if remaining the less attractive social media with fewer visitors' interactions. This will be a challenge until the end of the project, but the conclusion is that this social media does not work as well as it did in the past, mostly because of the structural changes it faced. So, regarding the powerful results within these networks, one can conclude that LinkedIn is the platform that works best for the project.

Activity	Results
Website	4.641 visits
LinkedIn	85 posts; 44.400 impressions; 980 reactions; 12 comments; 23 reposts
X	71 posts; 5 reposts; 4.442 views; 53 publications repost; 168 likes
YouTube	4 videos; 1.448 views
Newsletters	3 editions; 194 recipients
Brochures	2 editions; 150 printouts

TABLE 15 - COMMUNICATION RESULTS OF THE THIRD COMMUNICATION CAMPAIGN

Following all the results obtained from the third communication campaign, alongside the first two campaigns already executed, we can conclude that the efforts should be channelled to the events organisation and participation.

On the D5.3, a vulnerability was detected on the analytics indicator, stating that the website was the communication instrument that required more attention. Using a new analytics platform, one can conclude that the problem was not on the website, but on the metrics, confirming that the traffic on the website is, and has been, increasing. The mitigation actions planned to lead viewers to the website were executed anyway, also improving the website's performance.

Based on the results achieved so far, it can be concluded that the communication and dissemination activities are progressing as planned. The intention is to build on the work already carried out and to implement the proposed actions, ensuring that the results continue to develop in a positive direction.

3.2 NEXT STEPS

One more campaign is planned until the end of the project, starting in May 2026 and lasting until March 2027, with the purpose of promoting and exploiting the project's results. The targets at this stage are all audiences, with a special focus on specialised communities and operators, as well as end users, customers, and general citizens. Regardless, some core communications actions continue to be part of the project's strategy to disseminate the work it is developing.

A third Advisory Board Meeting is planned for 2026 to address the development of the project and discuss the final exploitation strategy and gather recommendations for post-project sustainability and impact. Building on the experience gained through the Horizon Booster and guided by recommendations from stakeholders and advisors, the final comprehensive documentation on exploitation pathways, business models, and sustainability mechanisms for all Key Exploitable Results (KERs) is being developed to finalize the overall exploitation strategy.

Regarding the dissemination boosters, until the end of this campaign, two more workshops are expected to happen, one regarding "RL applied to complex networks" in October 2026 and the other on "Co-learning and human-AI cooperation" in February 2027.

The planned workshop and policy maker meeting called "Ethics and regulation", co-organised with CLAIRE and led by the project partner ZHAW, is scheduled to happen in 2026, on a date yet to be determined.

The organisation of three AI open innovation competitions utilising the digital environments Grid2Op, BlueSky, and Flatland remains a key initiative to engage the global AI developer community. The three are planned to take place in Q3-Q4 of 2026. The Power Grid Domain Competition using the Grid2Op environment is led by RTE, the Railway Domain Competition is led by the Flatland Association, and the Air Traffic Management Competition, leveraging the BlueSky simulation platform, is led by TU Delft. Collectively, these competitions aim to attract international participation from researchers and domain experts, produce benchmark datasets and comparative insights, expand the contributor community for these digital ecosystems, and surface innovative approaches suitable for integration into future developments.

In addition to the planned activities of the project, a Winter School is being considered for implementation in Q4 2026 - Q1 2027 as part of the project's dissemination and exploitation activities. The event would be structured to provide researchers, students, and practitioners with comprehensive training on the application of AI4REALNET's digital environments, AI building blocks, and methodologies in the context of safety-critical infrastructure systems.

4. CONCLUSIONS

Considering the outcomes attained to date, the communication and dissemination efforts are moving in line with expectations. The focus going forward is to consolidate the work already undertaken and to carry out the planned activities, so that the results continue to improve and expand.

Therefore, 212 dissemination outcomes were achieved between months 19 and 30 of the project, divided between:

- Social media channels: 160 posts
- Newsletters: 3 sent
- Articles: 2
- Events: 33
- Scientific Publications issued: 13
- Scientific Publications in review: 1

The third communication campaign, implemented between April 2025 and April 2026, with the goal to reinforce the message of the technical solutions developed within the project's scope and enhance the project's results and best practices, was implemented with successful results, regarding the number of scientific publications issued in open-source, as well as the participation in more than 30 events.

So, one more communication campaign lies ahead until the end of the project, starting right after the submission of this deliverable, in April 2026.

Evidence of all the dissemination activities referred to in the deliverable is available in the annex section of this document.

REFERENCES

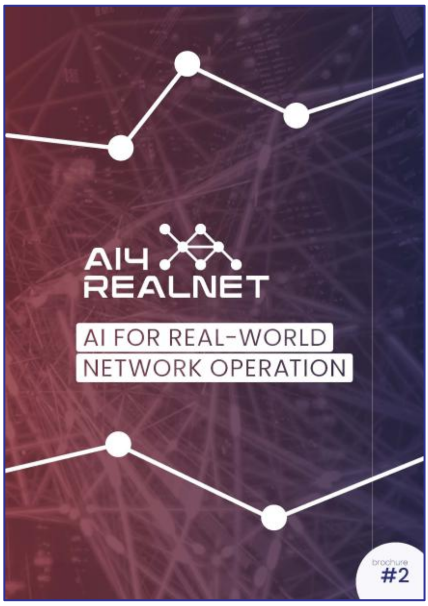
[1] AI4REALNET Grant Agreement number 101119527

[2] D5.1 – “Communication and dissemination plan”

[3] D5.3 – “Communication and dissemination monitoring phase 1”

[4] D6.1 – “Project management guide – procedures handbook”

ANNEX 1 – BROCHURE #2



AI4 REALNET

AI FOR REAL-WORLD NETWORK OPERATION

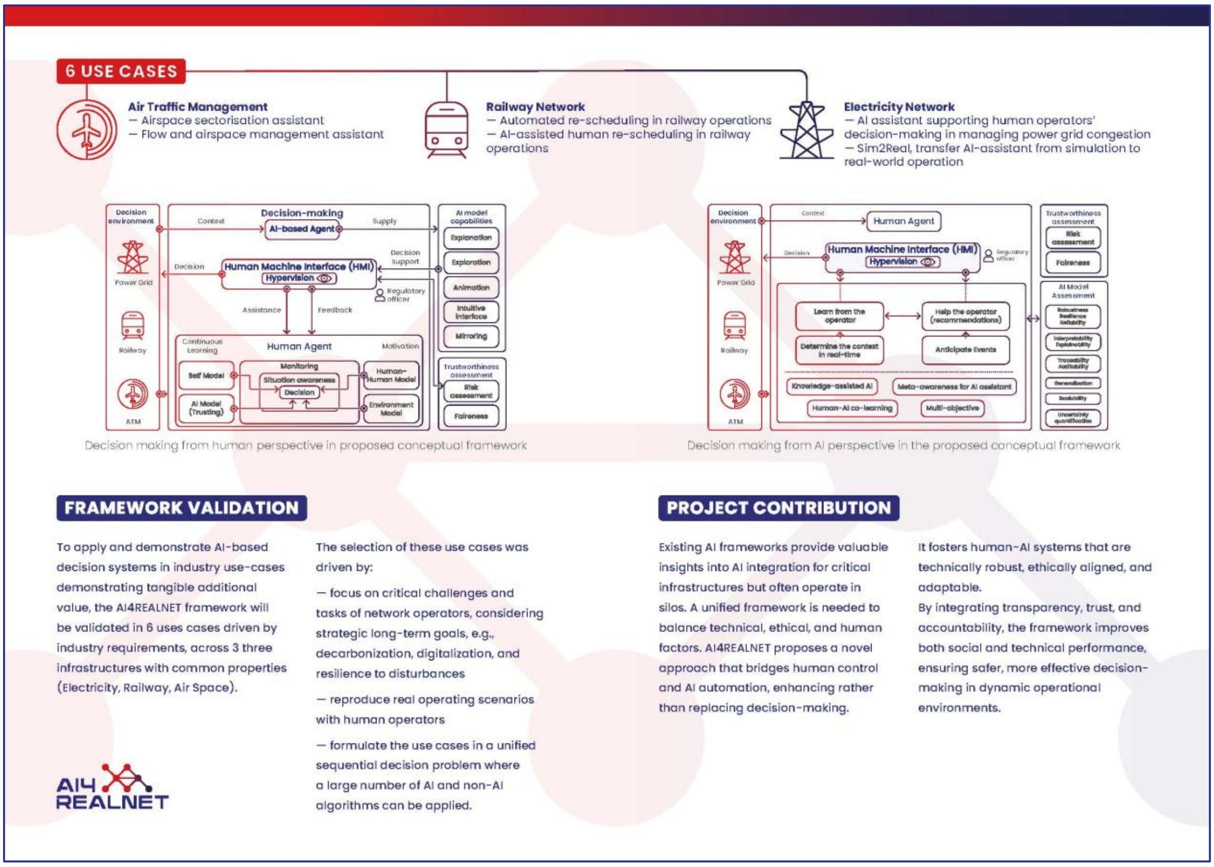
brochure #2

INESC TEC

ai4realnet.eu
ai4realnet@inesctec.pt
@ai4realnet-project
@AI4REALNET
@AI4REALNET
github.com/AI4REALNET

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

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



6 USE CASES

Air Traffic Management
— Airspace sectorisation assistant
— Flow and airspace management assistant

Railway Network
— Automated re-scheduling in railway operations
— AI-assisted human re-scheduling in railway operations

Electricity Network
— AI assistant supporting human operators' decision-making in managing power grid congestion
— Sim2Real, transfer AI-assistant from simulation to real-world operation

Decision making from human perspective in proposed conceptual framework

Decision environment
Power Grid
Railway
ATM
Context
Human Agent
Decision
Human Machine Interface (HMI)
Hypervision
Regulatory officer
Learn from the operator
Help the operator (recommendations)
Determine the context in real-time
Anticipate Events
Knowledge-assisted AI
Meta-awareness for AI assistant
Human-AI co-learning
Multi-objective
Trustworthiness assessment
Risk assessment
Fairness
AI Model Assessment
Robustness Resilience Reliability
Interpretability Explainability
Traceability Auditability
Generalization
Scalability
Uncertainty quantification

Decision making from AI perspective in the proposed conceptual framework

FRAMEWORK VALIDATION

To apply and demonstrate AI-based decision systems in industry use-cases demonstrating tangible additional value, the AI4REALNET framework will be validated in 6 uses cases driven by industry requirements, across 3 three infrastructures with common properties (Electricity, Railway, Air Space).

The selection of these use cases was driven by:

— focus on critical challenges and tasks of network operators, considering strategic long-term goals, e.g., decarbonization, digitalization, and resilience to disturbances

— reproduce real operating scenarios with human operators

— formulate the use cases in a unified sequential decision problem where a large number of AI and non-AI algorithms can be applied.

PROJECT CONTRIBUTION

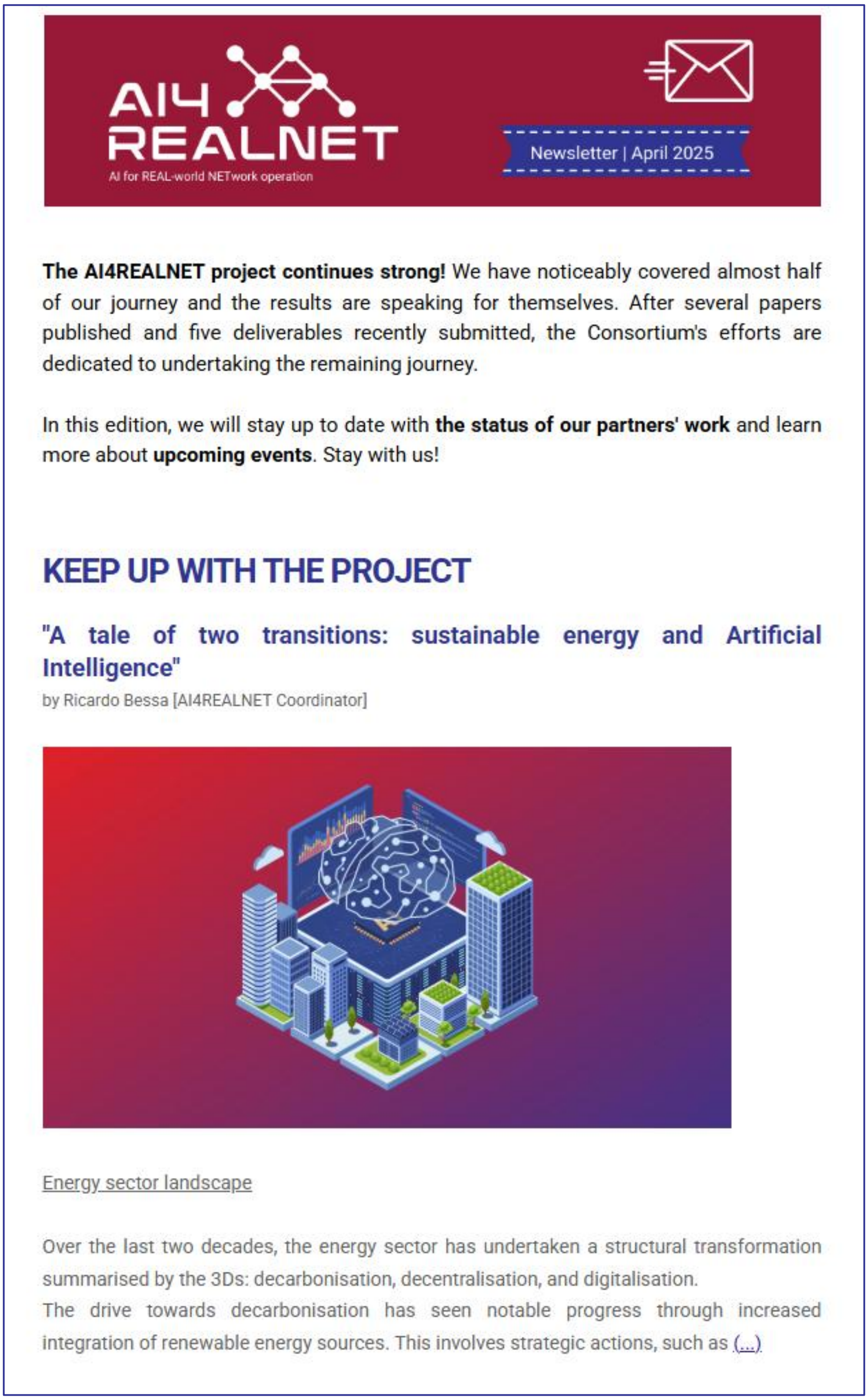
Existing AI frameworks provide valuable insights into AI integration for critical infrastructures but often operate in silos. A unified framework is needed to balance technical, ethical, and human factors. AI4REALNET proposes a novel approach that bridges human control and AI automation, enhancing rather than replacing decision-making.

It fosters human-AI systems that are technically robust, ethically aligned, and adaptable.

By integrating transparency, trust, and accountability, the framework improves both social and technical performance, ensuring safer, more effective decision-making in dynamic operational environments.

AI4 REALNET

ANNEX 2 – NEWSLETTER #5



AI4 REALNET

AI for REAL-world NETwork operation

Newsletter | April 2025

The AI4REALNET project continues strong! We have noticeably covered almost half of our journey and the results are speaking for themselves. After several papers published and five deliverables recently submitted, the Consortium's efforts are dedicated to undertaking the remaining journey.

In this edition, we will stay up to date with **the status of our partners' work** and learn more about **upcoming events**. Stay with us!

KEEP UP WITH THE PROJECT

"A tale of two transitions: sustainable energy and Artificial Intelligence"

by Ricardo Bessa [AI4REALNET Coordinator]

Energy sector landscape

Over the last two decades, the energy sector has undertaken a structural transformation summarised by the 3Ds: decarbonisation, decentralisation, and digitalisation.
The drive towards decarbonisation has seen notable progress through increased integration of renewable energy sources. This involves strategic actions, such as (...)

ANNEX 3 – ECML PKDD 2025 WORKSHOP ON CRITICAL INFRASTRUCTURES

The set of papers presented at the workshop shows that trustworthy AI for critical infrastructures is heading away from isolated algorithmic performance claims and towards system-level assurance, where safety, robustness, accountability, and operational feasibility must be engineered together. Across the contributions, a recurring theme emerges: critical infrastructures cannot afford to treat AI as a "smart component" inserted into an otherwise deterministic control chain. Instead, AI must be embedded as a managed and supervised element of a socio-technical system, continuously evaluated, constrained, and aligned with safety envelopes and human operational priorities.

A central contribution of the workshop papers is the explicit recognition that traditional correctness is insufficient for autonomous or AI-driven decision-making in critical infrastructures. Atif et al. argue that complex ML-based functions inherently face "unknown inputs" that cannot be exhaustively tested, thus undermining classical verification approaches and certification logic. Their position reframes the problem: infrastructures should not demand perfect correctness, but rather trustworthiness, supported through architectural patterns such as monitoring, redundancy, rejection mechanisms, and diversity (Atif et al., 2025). This is particularly relevant for infrastructures such as transportation networks, energy systems, or automated surveillance, where operational environments are open-ended, and the cost of failure is high. Their argument introduces a key challenge: infrastructures need engineering methods to make systems dependable even when components behave unpredictably, rather than attempting to eliminate unpredictability entirely (Atif et al., 2025). This is a major shift in how safety cases for AI-enabled infrastructure control can be constructed.

A second important contribution lies in the reinforcement learning (RL) domain, where several papers highlight both the promise and the fragility of RL for infrastructure control. King et al. explore a cruise ship HVAC optimization scenario in which RL is used to improve energy efficiency under uncertain environmental and operational conditions. Their work introduces an important link between classical hazard analysis and RL safety: they use System-Theoretic Process Analysis (STPA) to identify hazards and translate them into constraints and controller requirements, then implement these constraints and requirements as a safety shield that blocks unsafe actions (King et al., 2025). The study reveals a practical challenge: shielding improves safety during deployment, but blocking unsafe actions during training can hinder the RL agent's ability to learn a safe policy. This reflects a critical infrastructure dilemma: safety enforcement mechanisms may distort learning dynamics, creating a tension between exploration and constraint satisfaction that is rarely acknowledged in theoretical RL research. Their paper thus contributes an applied insight: safety must be treated not only as a deployment property, but also as a training-time design variable, particularly when infrastructures demand predictable behavior from the first day of operation (King et al., 2025).

This challenge is further reinforced by the broader perspective provided by (Yamagata et al., 2025), which addresses the fragmentation of terminology and methodology in safe RL. The authors propose consensus definitions and a structured checklist spanning ethical requirements, reward specification, robustness against uncertainty, explicit constraints, the use of simulators and expert knowledge, and finally, safety layers such as shielding and human intervention. For critical infrastructures, this paper's key contribution is not a new algorithm but rather a structured engineering viewpoint: safe RL cannot be reduced to constrained optimization alone. It must incorporate traceability, explainability, reward alignment, and mechanisms for human override. The paper also highlights a major infrastructure challenge: the lack of standardized guidelines makes it difficult for operators and regulators to judge whether an RL-based controller is deployable. In safety-critical domains, such ambiguity is itself a risk, because it blocks reproducibility and certification readiness.

The workshop also provides evidence that AI safety challenges in infrastructures are increasingly tied to security and adversarial resilience, not only physical failures. Lowe et al. contribute a strong example in the space domain, focusing on satellite object detection. Their work proposes a statistical method based on Mahalanobis distance to distinguish between adversarial perturbations and natural sensory anomalies. Their findings show that adversarial attacks can degrade detection performance at perturbation levels where natural noise does not trigger detection, and that common anomaly detection methods such as variational autoencoders may fail to detect adversarial manipulations (Lowe et al., 2025). This contribution is highly relevant to critical infrastructure because it highlights the necessity of threat differentiation: an infrastructure operator must respond differently to a sensor anomaly than to a deliberate adversarial manipulation. The paper frames adversarial detection not as an isolated ML security task but as part of a broader monitoring architecture supporting safe autonomous operation, which aligns closely with infrastructure supervisory control principles.

Another strong infrastructure-oriented contribution is the work by Schiendorfer et al., which addresses gas grid management through multi-objective reinforcement learning (MORL) and reports lessons learned from real-world pilot deployments. Their paper emphasizes that infrastructure control is inherently multi-criteria: operators must balance linepack, pressure limits, operational smoothness, and safety constraints, often under changing priorities. Their approach uses the GPI-PD algorithm to maintain a set of policies optimized for different objective weightings, allowing operators to select strategies depending on real-time needs (Schiendorfer et al., 2025). The paper's main practical insight is that deployment is limited not by the availability of RL algorithms but by infrastructure realities: the effort required for labeling, the difficulty of generating negative training examples, the scalability limitations of high-fidelity simulation, and the importance of interpretability and trust for dispatchers. Their lessons highlight a key challenge for critical infrastructure: AI deployment is fundamentally constrained by data availability, operational acceptance, and the ability to integrate AI into existing SCADA and decision-support workflows (Schiendorfer et al., 2025). This paper demonstrates that the true bottleneck is not "learning a policy," but rather building an ecosystem in which a policy can be trusted and adopted.

Complementing these technical works, the paper on continuous assessment-driven requirement elicitation (Fedorova et al., 2025) makes an important contribution to governance and engineering processes. It adapts the European Commission's ALTAI framework into a lifecycle method that distinguishes ethical requirements relevant at different Technology Readiness Levels (TRLs), enabling a staged approach to ethics-by-design. Their railway management use case shows how ALTAI questions can be converted into functional and non-functional requirements, while explicitly classifying which issues are relevant for proof-of-concept and which will become relevant in full-scale deployment. For critical infrastructures, this is particularly important because infrastructure AI systems are rarely deployed in a single step; they evolve from prototypes to operational systems, and ethical risks often shift across this maturity trajectory. Their contribution highlights a challenge that is often underestimated: assessment frameworks designed for ex-post evaluation may foster false confidence when used prematurely. The authors therefore argue that trustworthy AI requires continuous assessment, aligned with regulatory requirements such as the EU AI Act’s lifecycle risk management obligations. This paper reinforces the idea that trustworthiness is not solely a technical attribute but a development discipline, linking requirements engineering to compliance and operational readiness.

From these contributions, several promising research lines become evident. One major direction is the development of architectural trust patterns for AI components, extending beyond monitoring into compositional assurance. Atif et al.’s call to design for trust rather than correctness suggests a need for reusable architectural templates for critical infrastructure, including runtime monitors, redundancy mechanisms, reject options, and fail-safe strategies that are compatible with certification. This intersects directly with Lowe et al.’s work, where detection and supervision architectures become key. A future research line would therefore focus on unifying monitoring approaches for uncertainty, anomalies, and adversarial threats into a standardized supervisory layer for AI-enabled infrastructure systems. This could lead to reference architectures that integrate epistemic uncertainty estimation, distribution shift detection, adversarial attribution, and safe fallback policies, with explicit interfaces to human operators.

A second research line concerns safe reinforcement learning training methodologies that avoid the paradox highlighted by King et al.: shielding improves safety but may prevent the agent from learning. This suggests a need for adaptive shielding strategies, staged curricula, or dual-mode learning approaches where safety constraints are gradually tightened as policies mature. The checklist-based framework for safe RL suggests the importance of structured design workflows, but further research is needed to operationalize these workflows into toolchains that infrastructure operators can apply. The MORL pilot deployment in gas grids also suggests that RL must be embedded in decision-support paradigms rather than replacing dispatchers. Research could therefore explore interactive RL and MORL frameworks in which policy generation is automated but selection and accountability remain human-centered, aligning with “human-in-command” infrastructure principles.

A third research line involves data generation and simulation fidelity. Both gas grid and cruise ship papers show that realistic simulators are too slow or incomplete, while purely data-driven surrogate models have blind spots due to limited state coverage. A promising direction is hybrid simulation architectures combining physics-based models, surrogate neural regressors, and scenario generation engines that produce rare failure states. This is essential for infrastructures because historical data is biased toward safe operation, leaving insufficient negative examples for learning safe responses. Methods such as synthetic stress testing, counterfactual simulation, and adversarial scenario generation could become core components of infrastructure AI validation pipelines.

A fourth research line concerns requirements engineering and governance mechanisms that scale. The ALTAI adaptation paper demonstrates that trustworthiness frameworks must be lifecycle-sensitive. Future research could focus on formalizing mappings between ethical assessment checklists, infrastructure safety constraints, and verification artifacts such as safety cases. This could produce a bridge between EU regulatory compliance and technical design, enabling infrastructure operators to build auditable evidence chains linking requirements, training data decisions, algorithmic constraints, monitoring policies, and operational logs. Such work is crucial because infrastructures must demonstrate not only performance but also accountability, explainability, and controlled risk exposure.

The synergies between the papers are particularly strong and point to an emerging integrated vision. The concept of "unknown inputs" and trust-based architecture from Atif et al. aligns naturally with Lowe et al.'s work on distinguishing adversarial from natural anomalies, since both are fundamentally about detecting when the AI system is operating outside its validated regime. Their approaches could be unified: unknown-input handling could be enriched with threat classification mechanisms, enabling infrastructure systems to not only reject uncertain outputs but also infer whether the uncertainty is environmental or malicious. Similarly, the STPA-driven safety shielding approach aligns closely with the safe RL checklist paper: shielding is explicitly identified as a safety layer, and STPA provides a systematic method for deriving the constraints that shielding must enforce. This provides a bridge between hazard analysis and RL engineering, which is essential in infrastructure where safety constraints must be justified at the system level.

There is also a strong synergy between the gas grid MORL deployment and the ALTAI-driven requirement elicitation paper. The gas grid work highlights the practical difficulty of labeling and operator trust, while the ALTAI-based method provides a structured mechanism to elicit requirements for transparency, accountability, and fairness early in the design. Together, they suggest a holistic infrastructure workflow: ethical and trustworthiness requirements are elicited continuously across TRL stages, then translated into system constraints, monitoring requirements, logging policies, and decision support interfaces that operators can understand and accept. In this sense, MORL multi-policy approaches can serve as a practical implementation of "human agency and oversight," since operators can choose among policies rather than blindly executing a single learned policy.

Finally, the workshop shows that trustworthiness in critical infrastructures must be approached as a multi-layered ecosystem problem, where algorithmic advances alone are insufficient. Robustness to uncertainty, safety enforcement, monitoring and anomaly detection, multi-objective decision-making, and lifecycle requirement elicitation all form interdependent layers. The synergy is that each paper contributes a piece of a unified stack: requirements and governance (ALTAI-based elicitation), architectural trust patterns (Beyond Correctness), monitoring and adversarial differentiation (Mahalanobis-based detection), RL design guidelines (checklists and definitions), safety enforcement mechanisms (STPA-based shields), and operational deployment experience (gas grid MORL pilots). Taken together, they suggest a mature research agenda: critical infrastructures require AI systems that are not only accurate, but also continuously assessed, architected for uncertainty, safeguarded by explicit constraints, supervised against anomalies and attacks, and designed to preserve human authority.

Atif, M., Zoppi, T., & Bondavalli, A. (2025). Beyond correctness: Architecting trustworthy software for autonomous systems in the age of AI. ECML PKDD 2025, Porto, Portugal.

Fedorova, A., Stefani, W., Heitz, C., Chavarriaga, R. (2025). Continuous assessment-driven requirement elicitation for trustworthy AI systems. (2025). ECML PKDD 2025, Porto, Portugal.

Yamagata, T., Santos-Rodriguez, R. (2025). Guidelines for safe and robust reinforcement learning: From definitions to design. ECML PKDD 2025, Porto, Portugal.

King, A., Shahinas, E., & Atmojo, U. (2025). Constructing safety shields for reinforcement learning agents with system-theoretic process analysis. ECML PKDD 2025, Porto, Portugal.

Lowe, R., Ulan, M., Bui, T. H., Giordana, G., Giani, T., & Rossi, F. (2025). Differentiating adversarial attacks from natural sensory anomalies in object detection. ECML PKDD 2025, Porto, Portugal.

Schiendorfer, A., Schmidt, T., Outafraout, K., Baschin, M., Hei, Y., Streubel, T., Kätzel, P., Michailov, L., Felix, R., & Harpeng, L. (2025). Multi-objective reinforcement learning for safety-critical gas grid management: Lessons from real-world pilot deployment. WECML PKDD 2025, Porto, Portugal.

ANNEX 4 – IHIET 2025 WORKSHOP ON HUMAN-AI COLLABORATION IN CRITICAL TASKS

The AI4REALNET-Project organised a session on "Human-AI Collaboration in Critical Tasks" at the 15th International Conference on Human Interaction & Emerging Technologies (IHIET 2025) at the University of Vienna, Vienna, Austria, August 25-27, 2025. The session was chaired by Toni Waefler (FHNW, Switzerland) and Jonas Lundberg (LiU, Sweden).

AI4 REALNET
SIG: Human-AI Collaboration in Critical Tasks
15th International Conference on Human Interaction & Emerging Technologies (IHIET; 26 August 2025)
Toni Wäfler (University of Applied Sciences and Arts Northwestern Switzerland)
Jonas Lundberg (Linköping University, Sweden)

Five members of the AI4REALNET consortium gave presentations:

Sebastiaan De Peuter, UvA, Netherlands: AI-Advised Decision Making in Critical Systems

In recent decades, AI has moved from outplaying humans in chess to solving intractable scientific problems such as protein folding. However, AI successes have largely been limited to well-defined problems. Most real-world decision problems, in contrast, are ill-defined because their goals and constraints are hard to describe in mathematical form. For critical systems, human operation therefore remains the gold standard. Though optimal control of these systems can be ill-defined, human operators generally have a tacit understanding of the full goals and constraints. Despite the potential for better-than-human performance, handing control of these systems to AI, which lack this tacit knowledge, risks unacceptably unsafe behaviours when it optimises ill-defined goals under incompletely described constraints. A promising alternative then is to create teams of human and AI operators that collaboratively control the system. Keeping human operators in the loop, where they can ensure that the team's decisions align with their tacit goals and constraints, guarantees safe operation, while the AI can improve performance by augmenting and complementing the human operator's decisions. Over time, collaboration with the human operator allows the AI to learn these tacit goals and constraints, and to improve both the quality and safety of its behaviour. AI-advised decision making (AIAD) is an example of such a human-AI team. In AIAD, a team consisting of a human decision maker and an AI assistant collectively solves a decision problem. The human decision maker has certain goals or constraints for this decision problem, but these are tacit and cannot be communicated to the AI assistant. AIAD supports different types of interaction with the human decision maker or to give the assistant partial control over a critical system.

Samira Hamouche, FHNW, Switzerland: Applying Job Design Criteria for Effective Human-AI Collaboration[1]

Human-AI collaboration often underperforms due to a lack of motivation-supportive system design. Work design theory – specifically the Job Characteristics Model (JCM) – can guide the development and evaluation of AI systems. Qualitative evaluation anchors can translate core job design criteria into assessable aspects of AI-supported work. These anchors are developed through a theory-driven process that combines work design theory with recent literature on AI's impact on work characteristics. The goal is to foster intrinsically motivating and cognitively engaging human roles in AI collaboration, thereby enhancing overall human-AI team performance.

Jonas Lundberg, LiU, Sweden: Augmenting human operators with Artificial Intelligence in critical network systems operation

The presentation addresses how to describe critical episodes of interaction between human operators and autonomous, automated, and manual control systems. It addresses three questions: (1) what levels of cognitive control are important to include in a descriptive framework for joint human-autonomy in process control; (2) how should one describe temporal developments in joint socio-technical systems; and (3) how does one analyse communication and control at the system joints. A framework for description and analysis is presented, the Joint Control Framework (JCF). It allows descriptions of the three previously mentioned aspects through three analytical activities: process mapping (PM), analysis of Levels of Autonomy in Cognitive Control (LACC), and temporal descriptions of human–machine interaction (T-HMI). This facilitates analyses across cases and domains. The framework is discussed in the presentation.

Jan Viebahn, Tennet, Netherlands: User experience evaluation of an AI-based decision-support tool for power grid congestion management[2]

The electricity system is changing rapidly, due to the increasing efforts against climate change. In the control room, power grid operators are already being challenged by the changing system behaviour, and maintaining a high level of security of supply is expected to become even more challenging in the future. To cope with these challenges, new tools and functionalities, such as AI-based decision support tools (DSTs) are needed. Developers of future DSTs must consider not only technical aspects, but also whether new systems are usable by power system operators. This study presents a case study of user

[1] Hamouche, S., Dettling, N., Waefler, T. (2025). Applying Job Design Criteria for Effective Human-AI Collaboration. In: Tareq Z. Ahram and Renate Motschnig (eds) Human Interaction and Emerging Technologies (IHIET 2025). AHFE (2025) International Conference. AHFE Open Access, vol 197. AHFE International, USA. http://doi.org/10.54941/ahfe1006695

[2] Viebahn, J., Ayedh, A., Lundberg, J., Bång, M., Keijzers, J. (2025). User experience evaluation of an AI-based decision-support tool for power grid congestion management. In: Tareq Z. Ahram and Renate Motschnig (eds) Human Interaction and Emerging Technologies (IHIET 2025). AHFE (2025) International Conference. AHFE Open Access, vol 197. AHFE International, USA. http://doi.org/10.54941/ahfe1006694

experience (UX) evaluation applied to a DST for power grid congestion management. The evaluation approach employs a broad range of UX metrics. More precisely, we (i) introduce entirely new UX metrics based on a cognitive analysis of the human-AI interactions, (ii) provide a questionnaire and a set of tasks that are tailor-made for the DST to assess acceptance, trust, and performance, and (iii) apply established generic questionnaires to assess usability and workload. At the same time, the employed methods are mostly simple, such that the evaluation requires relatively low effort. The complete end-user population participated in the study, and the DST exhibits high scores in almost all UX metrics. The results form a baseline of summative user research, which enables benchmarking of future congestion management tools (or future releases of the same tool).

Vitor Minhoto, INESC TEC, Portugal: Human system symbiosis – the next generation of human-AI implicit interaction

AI4REALNET European project is strongly focused on the creation of different levels of human-AI collaboration, from full human to full AI-based control. Independent of these levels, this project aims to have a human-centred AI approach to take advantage as much as possible of the knowledge, experience, and decision-making capabilities of humans. As a key aspect to bring humans even closer to AI is to try to understand the human psychological status – a system that adapts to the human mental state and reactivity increases empathy between humans and AI. We show how AI4REALNET is tackling this problem and what methodologies are being developed and envisioned to be used on the project and tested with human operators in simulated environments.

Learnings from the discussions with the audience of the session

The session exhibited the variety of research happening within AI4REALNET, with a focus on the interaction between human operators of critical systems and AI tools. The presented research covered a wide range of topics, ranging from understanding the human experience with these tools to AI capabilities that can improve that experience. All presentations highlighted the importance of user-centric design of AI tools, rather than the AI-centered design that is still all too common. Overall, the session reinforced a key message: the future of AI in critical systems depends on how well we design for synergy between human intuition and machine intelligence.

Human-AI Collaboration in critical tasks is a relevant theme that gathered the interest of the audience. The discussion around this topic highlighted the importance of designing systems that do not replace, but rather augment human expertise.

The discussions revealed that the AI4REALNET project is unique and compelling because it focuses on different modes of human-AI collaboration, ranging from AI assisting humans, co-learning between AI and humans, autonomous AI directed by humans, and exploring ways of AI to sense and respond to the human psychological state. These approaches make clear that designing AI support for critical network control is not just a technical challenge, but a complex endeavour of sociotechnical integration of humans, AI design, and process design.

IHIET
INTERNATIONAL
Austria
IHIET
International
International Conference on
Human Interaction & Emerging Technologies (IHIET)
WWW.IHIET.ORG

ANNEX 5 – ADVISORY BOARD MEETING

List of Attendees from the Advisory Board, July 2, 2025

Entity	Type	Sector	Representative name
ENTSO-E	Association	Energy	Ilaria Federici
Intel	Industry	Other	Valerio Frascolla
SSENSEI Advisory	Start-up	AI	Yves Lostanlen
DFKI	Academic	AI	Philipp Slusallek
Eurocontrol	Association	ATM	Adam B. Tisza
Eurocontrol	Association	ATM	Dirk Schaefer
DIGI Mind Sphere	Start-up	AI	Viktor Miloshevski
Artelys	SME	AI	Nicolas Lair
Federal Office of Transport (BAV)	Other	Railway	Dominic Brunner

Agenda

Time	Topic
14:00-14:05	Current status of the AI4REALNET project
14:05-14:15	AI4REALNET conceptual framework
14:15-14:40	Feedback from the AB
14:40-14:55	Digital environments new features
14:55-15:20	Feedback from the AB: *What are the interoperability needs? How can we avoid the risk of being too vertically oriented to specific use cases? How to validate against real scenarios?*
15:20-15:35	WP2 developments: AI building blocks
15:35-16:00	Feedback from the AB: *How can we avoid the risk of being too vertically oriented to specific use cases? Potential for re-use and upscaling of the AI models in different scenarios and use cases?*
16:00-16:15	WP4 development: evaluation protocols for AI

16:15-16:40	Feedback from the AB: *Results could be of relevance to the broader community thinking about AI safety, as well as to regulatory bodies. How can we improve the dissemination of these ideas? how to define specific KPIs or aggregation of KPIs which could better express the "augmentation potential" achieved by AI on top of the Human baseline? How do we conduct the cost-benefit analysis for a final expected TRL 4-5?*
16:40-16:45	Exploitation Strategy of AI4REALNET
16:45-17:05	Feedback from the AB: Feedback about the exploitation paths and how they can be improved/implemented

Meeting minutes

Conceptual framework

- In the conceptual framework schematic, the trustworthiness assessment block appears to be linked only to the monitoring and the Regulatory Officer. However, it was clarified that this connection should be maintained throughout all stages of the decision-making process, ensuring continuous involvement of the Regulatory Officer. Moreover, we adopt a trustworthy-by-design approach, having identified a comprehensive set of requirements and key performance indicators (KPIs) from the very beginning of the project. It was also clarified that the right-hand side of the diagram ("decision-making from AI-based agent point of view") primarily focuses on the AI perspective, while the left-hand side ("decision-making from human agent point of view") reflects the human perspective. The essential properties remain represented in the schematic, though the overall visual clarity and layout could benefit from improvements.
- An important question concerns the implementation of co-learning in the operation of critical infrastructures—specifically, whether it functions in real-time or is limited to the training phase. It was clarified that co-learning is not intended for scenarios where decisions must be made within seconds. Instead, it is designed for operational contexts that allow sufficient time for collaborative decision-making and learning. In such cases, the co-learning process supports joint adaptation and improvement, rather than intervening in safety-critical decisions requiring immediate responses (real-time).
    - Another related question is whether the learning process is conducted in two distinct phases—first training, and then, assuming successful results, deployment into production. It was clarified that this approach depends on the available decision window, referred to in the conceptual framework as the "last time to decide". Human involvement can vary depending on this time frame. For instance, when sufficient time is available, humans can contribute by providing examples and actively participating in co-learning with the AI system.

  - One possible approach is to delegate routine tasks to AI systems, where trust in their performance has been established, while reserving human involvement for the most critical and high-stakes decisions.

Digital environments

- What is the scalability of the air traffic management use cases within digital environments, particularly considering the complexity of the scenarios? The underlying concept of these use cases is centered on strategic operators who are not subject to real-time pressure, allowing for more deliberate and scalable decision-making processes.
- It was recommended to clearly identify which critical tasks are addressed by the project within the use cases and digital environments. It was clarified that even decisions made close to real-time typically require some level of predictive analysis beforehand. As a result, there is always a defined "last time to decide" in each use case, during which the human operator can evaluate and compare different options before action is taken.
- AI4REALNET work on interoperability is clearly defined and serves as a solid foundation for further development.

AI building blocks (WP2)

- Should a formal method be used for problem decomposition? The current approach relies on statistical methods and is entirely domain-agnostic. However, a key challenge lies in ensuring, from a mathematical standpoint, that the decomposition process does not alter the fundamental representation of the system.
- Additional details about the algorithms are necessary to properly assess their replicability. Developing a truly agnostic solution may prove challenging. As a way forward, a concrete example or proof of concept could be provided by applying the approach to a different domain to demonstrate its generalizability.
- The graph neural solver is tailored to specific applications. However, projects like GridFM demonstrate that this technology can be extended to a wide range of use cases, showcasing its versatility and scalability.

Evaluation protocols (WP4)

- The current number of KPIs is quite extensive, and in some cases, fewer might be more effective. Having too many can make prioritization challenging. It may be beneficial to streamline and refine the list of KPIs as the project progresses to focus on the most impactful ones.
- In the cost-benefit analysis, it is important to clearly define who the benefits are intended for. It is recommended to conduct a stakeholder analysis to better understand the needs and perspectives of each relevant group and to tailor the analysis accordingly. CAPEX and OPEX may vary significantly depending on the stakeholder, highlighting the need for a role-specific evaluation.
- The fact that several KPI are based on standardization initiatives at ISO/IEC, it was appreciated.

Exploitation strategy of the project

- The strategy is well-structured and clearly defined.

- Suggestion of contacting
    - Mozilla Foundation: https://www.mozillafoundation.org/en/
    - EPRI Open Power AI Consortium for dissemination: https://msites.epri.com/opai
    - Big Data Value Association can also be relevant for dissemination: https://bdva.eu/
    - Contact Eurostack: https://eurostack.eu/
- How to keep the project alive when it ends? Community cannot be the right name.
- A key consideration is how to ensure the sustainability and continued impact of the project after its official conclusion. The term "community" may not accurately capture the long-term structure or mechanism needed for this purpose. It is recommended to define clear ownership, roles, and incentives to keep the initiative active and relevant beyond its formal lifecycle.

ANNEX 6 – CONTRIBUTIONS TO ADRA SRIDA

Big-ticket

Cost-effective integration of **renewables** on a local level and scaling to national (and cross-border) level, leveraging AI and data to **augment decision-making** across different **temporal and spatial scales** (e.g., optimising grid operations, improving system resilience, and accelerating the energy transition).

State of the art and state of the industry

Transmission and distribution system operators have been using expert systems and model-driven tools for decades for different tasks, such as network protection systems, voltage and congestion management, and state estimation. Machine and statistical learning techniques (e.g., artificial neural networks, linear additive models, tree-based methods) are used in energy demand and renewable energy forecasting for the next hours and days and reach market maturity.

Generation companies use machine and statistical learning for generation forecasting at different time scales that support trading decisions, and they combine sensor data with information collected from drones for asset inspection and maintenance.

Consumers receive energy efficiency tips extracted from data and non-intrusive monitoring. Demand-side management remains essentially a manual process.

The volume and diversity of data are increasing in the energy sector, with more data coming from sensors installed in the electrical grid and from grid users such as domestic consumers with smart meters and appliances, smart buildings, or distributed energy resources (storage, electric vehicles, photovoltaic). Furthermore, external data sources such as remote sensing (satellite images, sky cameras, drones) or textual data (e.g., social media) have started to be used to improve predictability and support the implementation of actions that increase grid resilience.

Applications and impact

In the energy sector, AI brings the following unique selling points for different applications:

- Real-time control and fast decision-making in normal and emergency scenarios of the electric power system. Example: managing the impact of extreme weather events; fault detection in energy assets.
- Modelling of complex energy systems where model-driven is impractical alone or with partial observability or under uncertainty (e.g., from renewable energy) or non-stationarity environments. Example: electrical network congestion management.
- Accelerate simulations by building data-driven models or via physics-informed machine learning. Example: power system dynamic simulation; renewable energy resource assessment; smart buildings comfort simulation.
- Explain complex systems or mathematical optimisation problems to decision-makers. Examples: cascading events in power grids; electricity market clearing.
- Capacity to forecast events (and uncertainty) by combining heterogeneous information sources. Examples: solar energy forecasting; electric vehicles forecasting.

Challenges

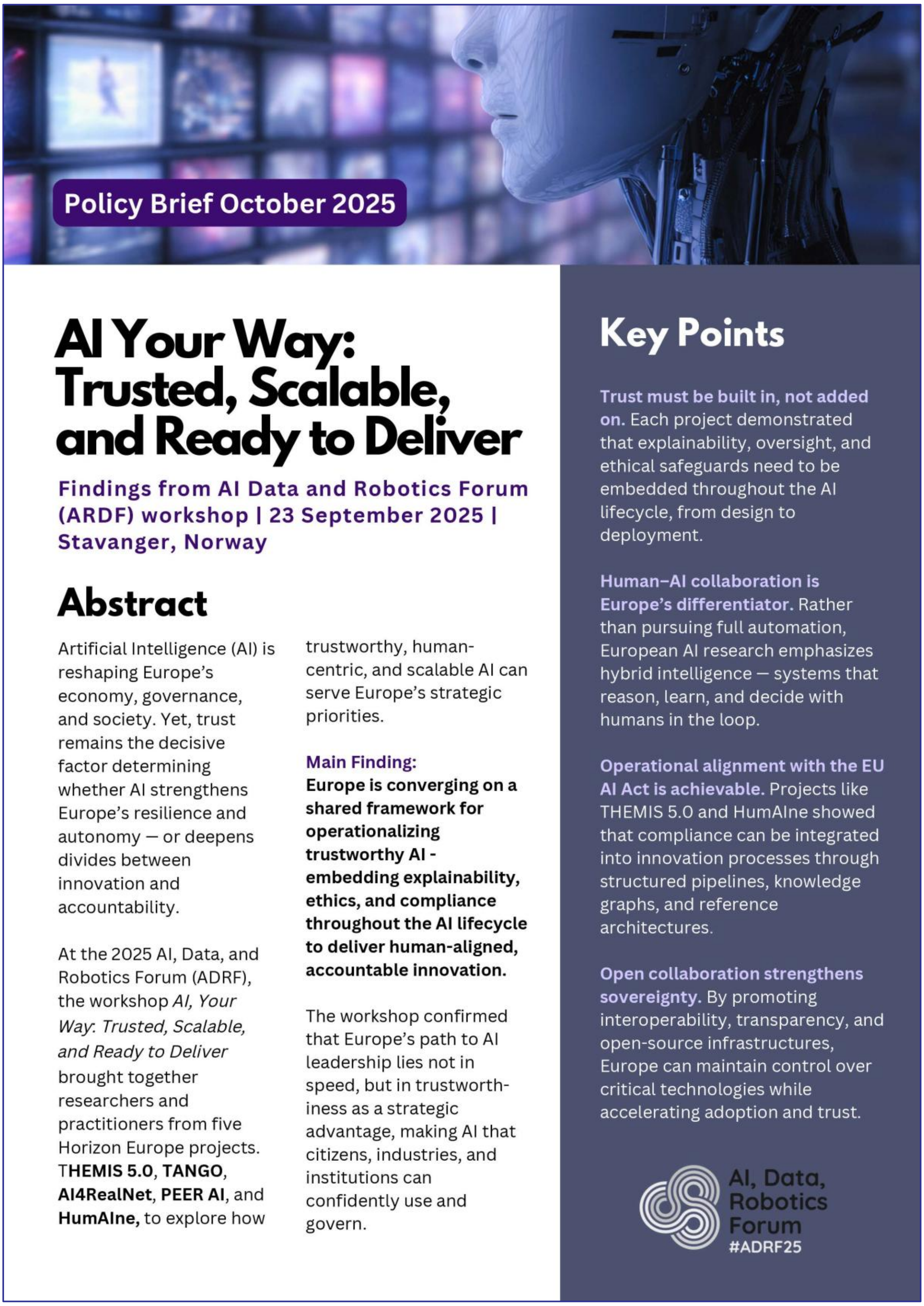
- Leverage decades of expert knowledge and model-driven approaches to advance AI applications in energy systems. This involves integrating deep domain expertise with data-driven and physics-informed models for applications like optimising grid operations, improving asset management, and energy market trading.
- AI-based systems should be more modular in the future and interoperable with other modules and systems.
- Data sharing is a major challenge in the energy sector. Furthermore, Data Spaces and Data Marketplaces can be key enablers of access to real and synthetic data, together with data semantic interoperability.
- Fast proof-of-concept methodologies are needed. This means agile construction of proof-of-concept demos for AI and sustainable open-source business models.
- Define and evaluate sustainable business models for energy Testing and Experimentation Facilities (TEF).

Roadmap

- Short-term:
  - Prioritize use cases in local energy communities to overcome data-sharing barriers in AI development and application. These communities are often more willing to share data in exchange for the benefits of data-driven services. Their use cases can serve as tangible examples of the value AI and data sharing bring, effectively demonstrating these benefits to larger energy utilities.
  - Continue advancing research in reinforcement learning, physics-informed machine learning, and neuro-symbolic learning, as these core algorithmic approaches are essential for modern AI systems. Their development aligns with the unique selling points of AI in the energy sector, as outlined in "Applications and Impact".
  - Work on problem decomposition for more scalable AI solutions and human interpretability in complex energy systems.
  - Develop AI assistants and human-AI co-learning schemes for tasks where humans still perform manual actions, e.g., power grid congestion management and energy market trading, predictive maintenance, and electrical network planning.
  - Use AI to support electrification of demand, e.g., optimal design of dynamic electricity tariffs (e.g., for electric vehicles), energy efficiency tips and investment assessment.
- Short-to-medium term:
  - Support and maximise the reuse of trained AI models without requiring a specific in-context relearning process and capable of combining heterogeneous data types. This should include AI foundation models tailored to use case scenarios and incorporate human feedback on preferences and reward functions, such as inverse reinforcement learning, to enhance adaptability and efficiency. This could support use cases like power grid congestion management, energy market trading, energy efficiency investment assessment.
  - Leverage AI to explain and extract knowledge and predict the spatial and/or temporal evolution of complex energy system operations (e.g., power grid, power plants). This should integrate physical assets and infrastructure, rule-based expert systems, model-

driven tools, and AI-based control software to enhance system understanding and decision-making.
    - Assess the performance, costs, and benefits of different architectures for distributed intelligence—centralised versus edge computing—across various use cases, such as power grids and local energy communities. This evaluation will support the selection of optimal AI deployment strategies for enhanced efficiency and scalability.
- Medium-to-long term:
    - Develop and deploy autonomous, self-healing AI-based systems for the operation of power grids, distributed energy resources and electricity market trading.
    - Promote the convergence of safety-critical systems and infrastructures (e.g., water and energy, energy and mobility) by aligning decision-making practices and AI system architectures. This integration will enhance resilience, efficiency, and interoperability across interconnected sectors.
    - Full interoperability between AI and legacy systems, and with human decision-makers.

ANNEX 7 – JOINT NEW POLICY BRIEF



Policy Brief October 2025

AI Your Way: Trusted, Scalable, and Ready to Deliver

Findings from AI Data and Robotics Forum (ARDF) workshop | 23 September 2025 | Stavanger, Norway

Abstract

Artificial Intelligence (AI) is reshaping Europe's economy, governance, and society. Yet, trust remains the decisive factor determining whether AI strengthens Europe's resilience and autonomy — or deepens divides between innovation and accountability.

At the 2025 AI, Data, and Robotics Forum (ADRF), the workshop *AI, Your Way: Trusted, Scalable, and Ready to Deliver* brought together researchers and practitioners from five Horizon Europe projects. **THEMIS 5.0**, **TANGO**, **AI4RealNet**, **PEER AI**, and **HumAIne,** to explore how trustworthy, human-centric, and scalable AI can serve Europe's strategic priorities.

Main Finding:
Europe is converging on a shared framework for operationalizing trustworthy AI - embedding explainability, ethics, and compliance throughout the AI lifecycle to deliver human-aligned, accountable innovation.

The workshop confirmed that Europe's path to AI leadership lies not in speed, but in trustworth-iness as a strategic advantage, making AI that citizens, industries, and institutions can confidently use and govern.

Key Points

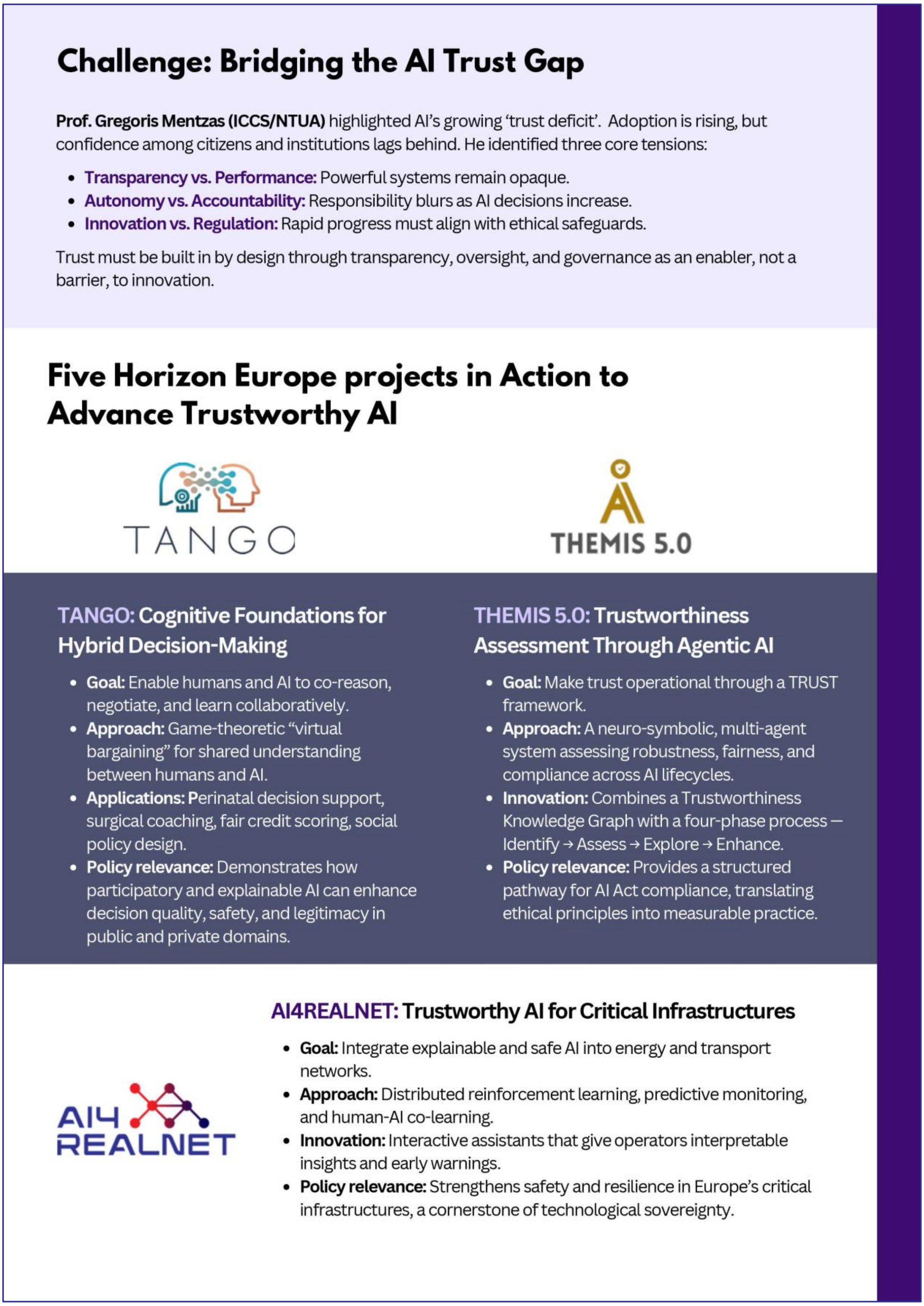
Trust must be built in, not added on. Each project demonstrated that explainability, oversight, and ethical safeguards need to be embedded throughout the AI lifecycle, from design to deployment.

Human–AI collaboration is Europe's differentiator. Rather than pursuing full automation, European AI research emphasizes hybrid intelligence — systems that reason, learn, and decide with humans in the loop.

Operational alignment with the EU AI Act is achievable. Projects like THEMIS 5.0 and HumAIne showed that compliance can be integrated into innovation processes through structured pipelines, knowledge graphs, and reference architectures.

Open collaboration strengthens sovereignty. By promoting interoperability, transparency, and open-source infrastructures, Europe can maintain control over critical technologies while accelerating adoption and trust.

AI, Data, Robotics Forum
#ADRF25



Challenge: Bridging the AI Trust Gap

Prof. Gregoris Mentzas (ICCS/NTUA) highlighted AI's growing 'trust deficit'. Adoption is rising, but confidence among citizens and institutions lags behind. He identified three core tensions:

- **Transparency vs. Performance:** Powerful systems remain opaque.
- **Autonomy vs. Accountability:** Responsibility blurs as AI decisions increase.
- **Innovation vs. Regulation:** Rapid progress must align with ethical safeguards.

Trust must be built in by design through transparency, oversight, and governance as an enabler, not a barrier, to innovation.

Five Horizon Europe projects in Action to Advance Trustworthy AI

TANGO

THEMIS 5.0

TANGO: Cognitive Foundations for Hybrid Decision-Making

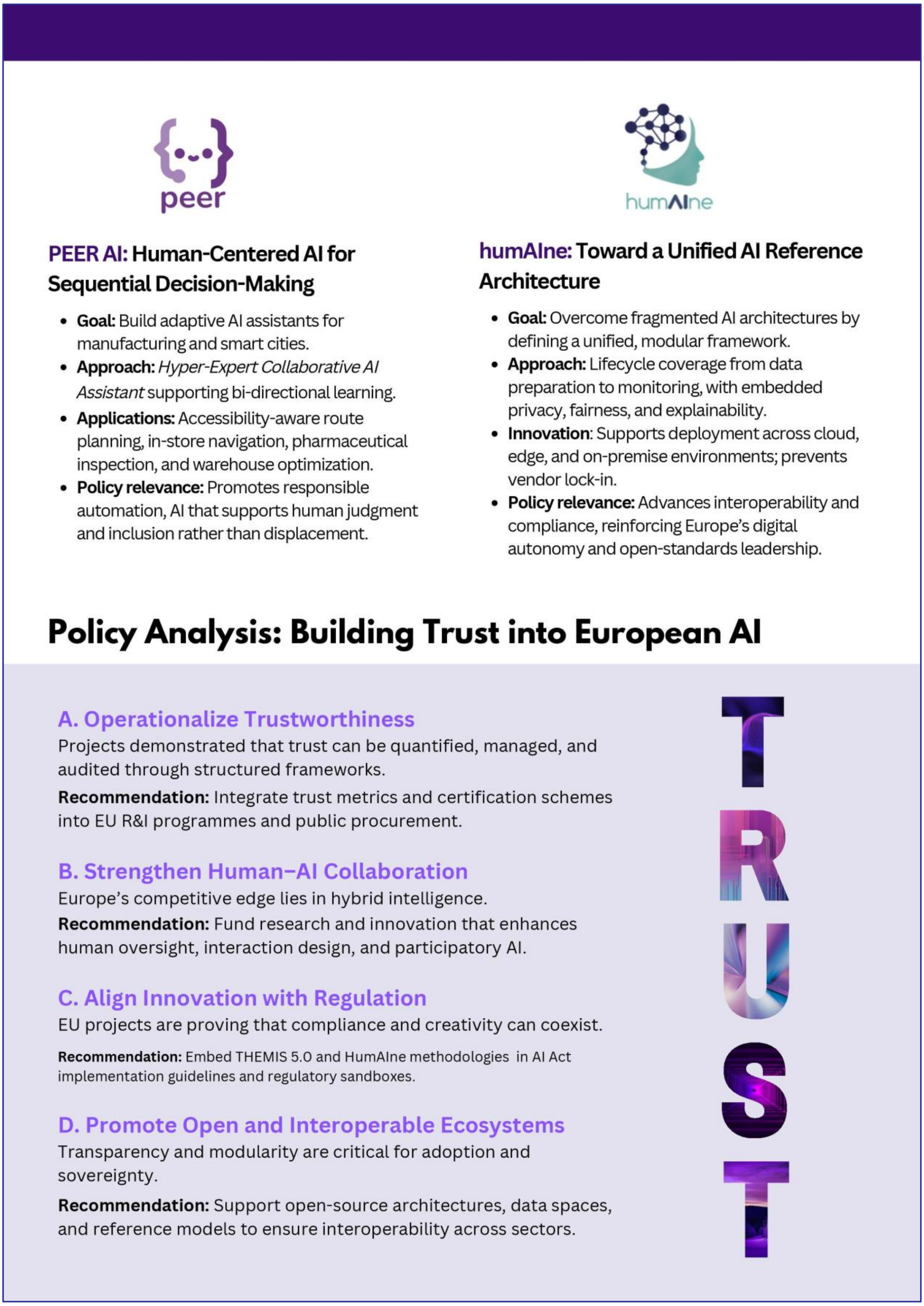
- **Goal:** Enable humans and AI to co-reason, negotiate, and learn collaboratively.
- **Approach:** Game-theoretic "virtual bargaining" for shared understanding between humans and AI.
- **Applications:** Perinatal decision support, surgical coaching, fair credit scoring, social policy design.
- **Policy relevance:** Demonstrates how participatory and explainable AI can enhance decision quality, safety, and legitimacy in public and private domains.

THEMIS 5.0: Trustworthiness Assessment Through Agentic AI

- **Goal:** Make trust operational through a TRUST framework.
- **Approach:** A neuro-symbolic, multi-agent system assessing robustness, fairness, and compliance across AI lifecycles.
- **Innovation:** Combines a Trustworthiness Knowledge Graph with a four-phase process — Identify → Assess → Explore → Enhance.
- **Policy relevance:** Provides a structured pathway for AI Act compliance, translating ethical principles into measurable practice.

AI4REALNET: Trustworthy AI for Critical Infrastructures

- **Goal:** Integrate explainable and safe AI into energy and transport networks.
- **Approach:** Distributed reinforcement learning, predictive monitoring, and human-AI co-learning.
- **Innovation:** Interactive assistants that give operators interpretable insights and early warnings.
- **Policy relevance:** Strengthens safety and resilience in Europe's critical infrastructures, a cornerstone of technological sovereignty.



PEER AI: Human-Centered AI for Sequential Decision-Making

- **Goal:** Build adaptive AI assistants for manufacturing and smart cities.
- **Approach:** *Hyper-Expert Collaborative AI Assistant* supporting bi-directional learning.
- **Applications:** Accessibility-aware route planning, in-store navigation, pharmaceutical inspection, and warehouse optimization.
- **Policy relevance:** Promotes responsible automation, AI that supports human judgment and inclusion rather than displacement.

humAIne: Toward a Unified AI Reference Architecture

- **Goal:** Overcome fragmented AI architectures by defining a unified, modular framework.
- **Approach:** Lifecycle coverage from data preparation to monitoring, with embedded privacy, fairness, and explainability.
- **Innovation**: Supports deployment across cloud, edge, and on-premise environments; prevents vendor lock-in.
- **Policy relevance:** Advances interoperability and compliance, reinforcing Europe's digital autonomy and open-standards leadership.

Policy Analysis: Building Trust into European AI

A. Operationalize Trustworthiness

Projects demonstrated that trust can be quantified, managed, and audited through structured frameworks.

Recommendation: Integrate trust metrics and certification schemes into EU R&I programmes and public procurement.

B. Strengthen Human–AI Collaboration

Europe's competitive edge lies in hybrid intelligence.

Recommendation: Fund research and innovation that enhances human oversight, interaction design, and participatory AI.

C. Align Innovation with Regulation

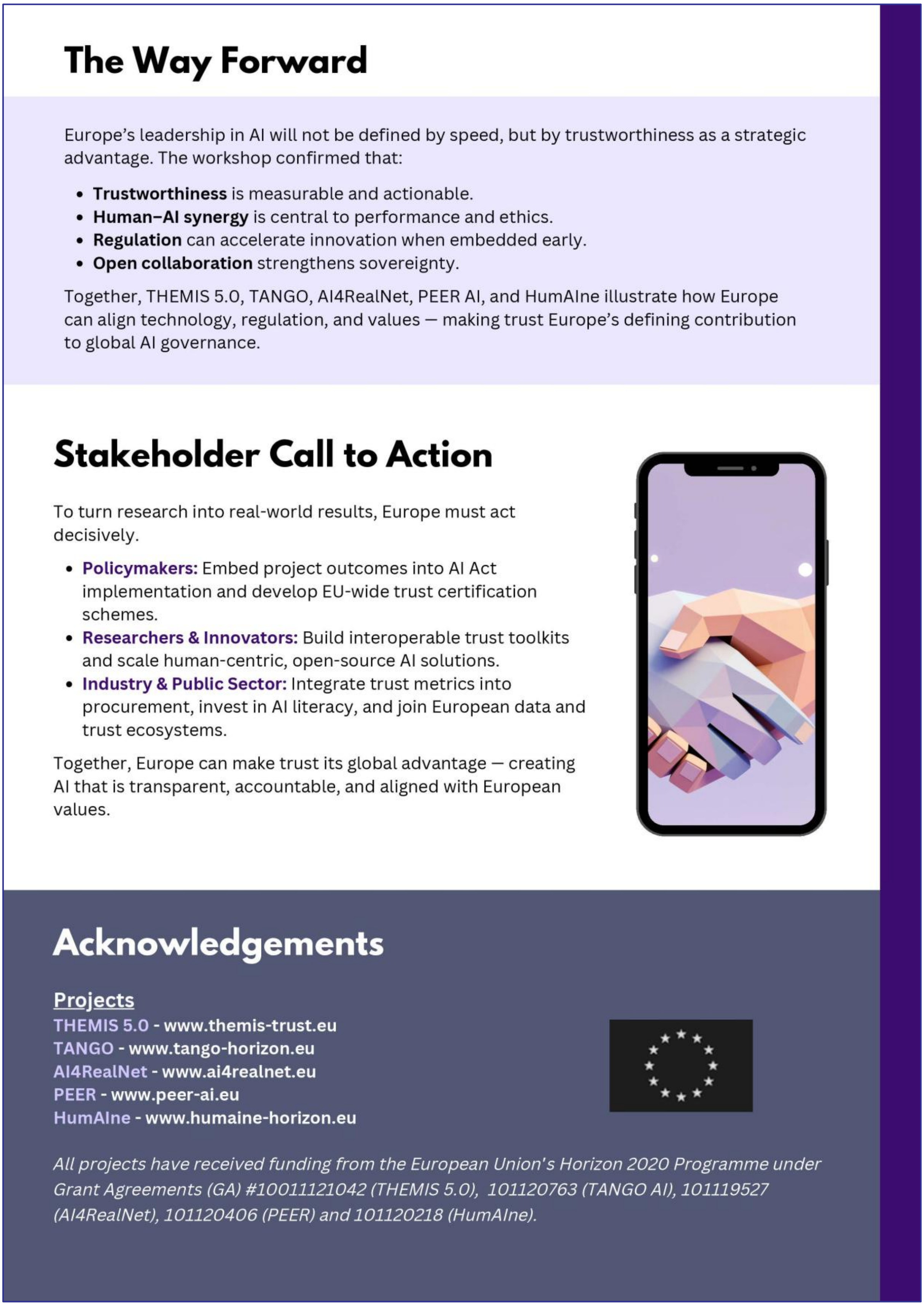
EU projects are proving that compliance and creativity can coexist.

Recommendation: Embed THEMIS 5.0 and HumAIne methodologies in AI Act implementation guidelines and regulatory sandboxes.

D. Promote Open and Interoperable Ecosystems

Transparency and modularity are critical for adoption and sovereignty.

Recommendation: Support open-source architectures, data spaces, and reference models to ensure interoperability across sectors.



The Way Forward

Europe's leadership in AI will not be defined by speed, but by trustworthiness as a strategic advantage. The workshop confirmed that:

- **Trustworthiness** is measurable and actionable.
- **Human–AI synergy** is central to performance and ethics.
- **Regulation** can accelerate innovation when embedded early.
- **Open collaboration** strengthens sovereignty.

Together, THEMIS 5.0, TANGO, AI4RealNet, PEER AI, and HumAIne illustrate how Europe can align technology, regulation, and values — making trust Europe's defining contribution to global AI governance.

Stakeholder Call to Action

To turn research into real-world results, Europe must act decisively.

- **Policymakers:** Embed project outcomes into AI Act implementation and develop EU-wide trust certification schemes.
- **Researchers & Innovators:** Build interoperable trust toolkits and scale human-centric, open-source AI solutions.
- **Industry & Public Sector:** Integrate trust metrics into procurement, invest in AI literacy, and join European data and trust ecosystems.

Together, Europe can make trust its global advantage — creating AI that is transparent, accountable, and aligned with European values.

Acknowledgements

Projects
THEMIS 5.0 - **www.themis-trust.eu**
TANGO - **www.tango-horizon.eu**
AI4RealNet - **www.ai4realnet.eu**
PEER - **www.peer-ai.eu**
HumAIne - **www.humaine-horizon.eu**

All projects have received funding from the European Union's Horizon 2020 Programme under Grant Agreements (GA) #10011121042 (THEMIS 5.0), 101120763 (TANGO AI), 101119527 (AI4RealNet), 101120406 (PEER) and 101120218 (HumAIne).

ANNEX 8 – HACK4RAIL 2025 CHALLENGE

🚄 Human–AI Teaming in Railway Operations: Learning Through Simulation

Daniel Boos[1], Adrian Egli[1], Roman Liessner[2], Manuel Meyer[3], Manuel Schneider[3]

[1] Swiss Federal Railways (SBB) - [2] DB InfraGO AG - [3] Flatland Association

As part of Hack4Rail 2025, we organized a unique challenge:

Development of an explainable human-AI collaboration interface for railway traffic management (Flatland).

As rail networks become increasingly complex and dense, the need for intelligent systems that support real-time dispatch and rescheduling grows. However, success depends not only on technical performance but also on how well the human-AI collaboration is designed.

A working prototype was developed using the Flatland simulation environment.

The main goal, however, was not to provide a perfect system, but to create a learning space: Participants were encouraged to engage with the topic, explore initial implementation ideas, and understand why human-AI teaming (HAT) is critical to the future of rail operations.

The prototype simulates a small but typical disruption scenario: A train stops unscheduled, which means that all other trains will have to be rerouted, reprioritized or rescheduled. Participants explored how a "smart agent" could support dispatchers in solving the problem. They integrated a simple algorithm that generates suggestions and visualizes them so dispatchers can quickly identify necessary changes.

A chatbot was also integrated to enable dialog-based interaction: Dispatchers can ask questions or give instructions in natural language, to which the system responds with relevant details—for example, the advantages and disadvantages of the proposed solution. The goal was to test how humans and AI act as interdependent, coordinated units and how a meaningful dialogue between humans and machines can emerge.

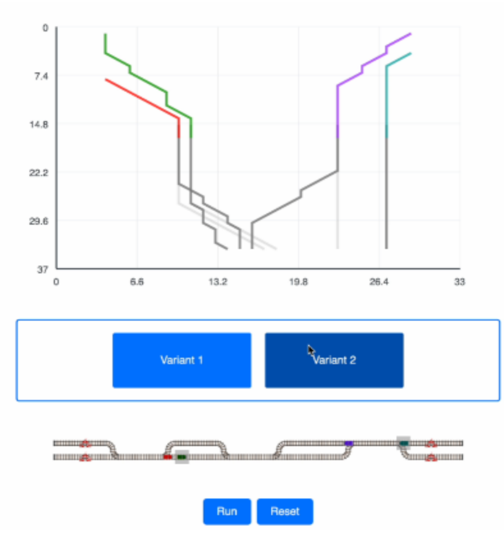
What participants learnt.

In the future, machines will no longer act merely as tools, but as active teammates. Participants agreed that the design of this interaction holds enormous potential. A key question was whether AI can make autonomous decisions based on observations – or whether humans need a system that quickly presents alternative options for them to choose from. The discussions were intense and inspiring.

Particularly impressive was the participants' intensive engagement with the topic, the many questions, and the many ideas. There was agreement that the insights gained will be valuable for daily work and future projects. With the increasing prevalence of AI-based applications in the rail sector, this topic will gain in importance. Especially when large AI systems are used in real-world workflows and can make decisions on their own, their success strongly depends on how well the collaboration between humans and machines is designed.

Results

The prototype has been published as an open-source project and serves as the basis for further research and development:



🔗 flatland-association/flatland-hmi

This project demonstrates that **human-AI teaming is more than just technology – it is a new way of thinking** about collaborative decision-making. The first step begins with curiosity, openness, and the courage to learn together.