



CALL FOR PROPOSALS
EXCHANGE SCHEMES
CATALOGUE OF CHALLENGES

***ENFIELD: EUROPEAN LIGHTHOUSE TO MANIFEST
TRUSTWORTHY AND GREEN AI***



Co-funded by
the European Union

TABLE OF CONTENTS

Research Pillar: Green AI	4
G-AI.1 Technological Symbioses and Rebound Effects with AI	4
G-AI.2 Green AI Indexing for Efficient Image and Point Cloud Retrieval	5
G-AI.3 Energy-Efficient Multi-Agent AI for Code Generation in Digital Twin Understanding	6
G-AI.4 How Green is Physics-Informed Machine Learning?	7
Research Pillar: Adaptive AI	8
A-AI.1 AI-Enhanced Visual Understanding for Inspection and Maintenance	8
A-AI.2 Efficient Continual Learning Systems	10
A-AI.3 Lifelong Learning and Fine-tuning in LLMs	12
A-AI.4 Flexible plastics sorting system based on the combination of multiple vision sensors and adaptive AI	13
A-AI-IDM.5 Batch pattern mining in industrial signals	15
Research Pillar: Human-Centric AI	16
HC-AI.1 Empowering Human-AI Decision-Making through Real-Time User Modeling	16
HC-AI.2 Interpretability and uncertainty in predictive models	18
HC-AI.3 Human-Centric Explainable AI for Critical Digital Infrastructure	20
HC-AI.4 Human-Centered Causal Machine Learning for Real-World Decision Support	22
HC-AI.5 Explainable AI Methods for Human-AI Decision Making	24
Research Pillar: Trustworthy AI	26
T-AI.1 Security and Robustness of AI systems	26
T-AI.2 Privacy and Compliance of AI systems	27
T-AI.3 AI in Distributed systems	28
T-AI.4 Enhancing User Trust in AI-powered Chatbots using UBM	30
T-AI.5 Trustworthy ML based scheduling for the energy domain	32
T-AI.6 Unified approach to assess safety and security of AI systems	33
T-AI.7 LLM Safety and Security	35
Industrial Domain: Energy	37
IDE.1 AI for Green Energy curtailment minimization	37
IDE.2 Explainable fault detection and classification in power grids	38
IDE.3 Combine AI with LLM for clear human interaction with complex data	39
IDE.4 LLM based data queries and visualization	41
Industrial Domain: Manufacturing	43
IDM.1 Adversarial Robustness Metrics and Evaluation for AI-Driven Safety Systems	43
IDM.2 Bias-aware People Detection: Quantifying and Mitigating Dataset and Model Biases Towards Safety Certification	45
IDM.3 Explainable Uncertainty-Aware Decision-Making for AI Systems in Manufacturing	47
IDM-A.AI.4 Multimodal AI for Human-State Monitoring	48
IDM-A.AI.5 Egocentric perception for shopfloor operators	50
Industrial Domain: Space	52
IDS.1 Detection of potential water illegal abstractions using Artificial Intelligence and Earth Observation	52
IDS.2 Causal Machine Learning model to identify agricultural practices aiding in yield productivity improvement using Earth Observation (EO) data	53
IDS.3 Advancing Eelgrass Monitoring Using Stationary High-Altitude Balloons and Hyperspectral AI-based Earth Observations	54
IDS.4 Enhancing natural disaster response through satellite-based earth observation and language models	55

INTRODUCTION

This is the catalogue of research challenges for the fourth out of four open calls, under the ENFIELD¹ (European Lighthouse to Manifest Trustworthy and Green AI) project, co-funded by the European Union, to call for the collaboration of individual researchers. Through the ENFIELD Exchange Scheme open calls and the Financial Support to Third Parties (FSTP) mechanism, the project aims to attract the top-level researchers to conduct foundational research activities related to the proposed scientific/technological challenges in artificial intelligence, contributing to ENFIELD network creation and expansion to European AI labs.

¹ Grant Agreement n° 101120657, funded by the European Union.

Research Pillar: Green AI

G-AI.1 Technological Symbioses and Rebound Effects with AI
Keywords: Lifecycle Assessment (LCA), Rebound Effect, Technological Symbiosis
STATE-OF-THE-ART
AI applications present environmental risks, particularly with respect to generative AI, but some applications may also have environmental benefits, as analysed by the Climate Change AI community (Rolnick, 2022). Full lifecycle assessments of AI applications tend to show that benefits overcome risks, for instance in agricultural robotics (La Rocca, 2024) and building automation (Bracquene, 2020). These conclusions, however, are only valid if we observe no rebound effect in the adoption of AI technologies, which is a strong and unrealistic assumption. Luccioni, Strubell and Crawford identified at least 12 rebound effects that may apply to AI (Luccioni, 2025).
SCIENTIFIC CHALLENGES
The objective of the exchange would be to develop a method to identify and anticipate rebound effects related to AI applications. The main hypothesis taken here is that rebound effects occur when there is a symbiosis between two technologies or more, i.e. a positive interaction between the two technologies acting as enabler of each other. Symbioses can be captured with causal loop diagrams (Charpenay, 2024). In this context, the scientific challenge is to identify key technologies related to AI in various application domains, e.g. based on inventory analysis, required for lifecycle assessments. Research activities will start with a use case in manufacturing in cooperation with the Danish Technology Institute (DTI). A lifecycle assessment has been made for this use case, in which an AI-based solution is compared to a manual solution in a production line.
RESEARCH ACTIVITIES
<ul style="list-style-type: none"> • Review of the inventory of published lifecycle assessments of AI systems • Review of envisioned AI applications in ENFIELD industrial domains
EXPECTED RESULTS
One (1) review paper describing identified technological symbioses related to AI.
POSSIBLE HOST ORGANISATIONS / SUPERVISORS
IMT - Mines Saint-Étienne / Victor Charpenay

Research Pillar: Green AI

G-AI.2 | Green AI Indexing for Efficient Image and Point Cloud Retrieval

Keywords: Green AI, Efficient Data Indexing, Image Retrieval, 3D Point Clouds, Sustainable Computing

STATE-OF-THE-ART

Current methods for searching through large collections of images and 3D point clouds rely on computationally intensive AI models that are run for every new search. This process is slow and consumes significant energy. While scene graphs—which map out objects and their relationships in a scene—are a powerful concept, they are not yet effectively used to pre-index data for faster, more energy-efficient retrieval. A notable gap exists in a streamlined method for converting text-based search queries directly into these structured scene graph representations without costly intermediate steps.

SCIENTIFIC CHALLENGES

The central challenge is to develop an energy-efficient indexing system to accelerate the retrieval of images and 3D point clouds. This involves several key difficulties:

- Development of a compact yet descriptive representation for rapid indexing and lookup.
- Balancing trade-offs between the granularity and completeness representations and the computational efficiency during retrieval.
- Generalization across diverse datasets. Test and refine the developed indexing methodology across various datasets to ensure broad applicability. Datasets include 3D point cloud data from overhead power lines and remote sensing.
- Direct conversion of text-based queries

RESEARCH ACTIVITIES

- A validated, energy-efficient framework for indexing and retrieving images and point clouds that significantly outperforms traditional methods in speed and sustainability.
- A specialized Large Language Model (LLM) capable of directly converting textual queries into structured scene graph representations.
- Dissemination of the project results through at least one peer-reviewed journal article, with a preferred target being the Applied Artificial Intelligence journal.
- The release of the developed tools and frameworks as open-source software encourages further research and adoption by the community.

EXPECTED RESULTS

- A validated, energy-efficient framework for indexing and retrieving images and 3D point cloud that significantly outperforms traditional methods in speed and sustainability. Addresses use cases in the energy, space, and manufacturing verticals of ENFIELD.
- A specialized Large Language Model (LLM) capable of directly converting textual queries into structured scene graph representations.
- Dissemination of the project results through at least one peer-reviewed journal article, with a preferred target being the Applied Artificial Intelligence journal.
- The release of the developed tools and frameworks as open-source software encourages further research and adoption by the community.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[SINTEF](#) / Sagar Sen
[TELENOR](#) / Jeriek Van den Abeele

Research Pillar: Green AI

G-AI.3 | Energy-Efficient Multi-Agent AI for Code Generation in Digital Twin Understanding

Keywords: Green AI, Multiagent AI, AI for Software Engineering, Digital Twins

STATE-OF-THE-ART

State-of-the-art digital twins replicate the structure and behavior of complex systems, such as power grids and manufacturing value chains. Assets and their relationships are modeled as knowledge graphs, while their dynamic behavior is captured through multimodal time series data from sensors. Recent advances integrate code-generating AI agents, powered by language models, to query and extract information from digital twins. Complementary agents verify the correctness of the generated code in a feedback loop, enabling iterative refinement. This multi-agent approach enhances digital twin comprehension. A key challenge is achieving this capability efficiently, with minimal energy consumption, aligning with goals of sustainable, Green AI.

SCIENTIFIC CHALLENGES

The central challenge is to design an energy-efficient multi-agent AI system capable of interacting with digital twins to answer complex queries. This involves several key objectives:

- Develop a multi-agent AI framework leveraging open-source models and external libraries and tools
- Generate optimized code to extract answers from digital twins representing complex systems.
- Produce code that enables visualizations and interpretable outputs based on the digital twin’s data.
- Validate the approach across diverse industry verticals, including energy, manufacturing, and space.

RESEARCH ACTIVITIES

- Enhance SINTEF’s multi-agent AI system, LUMEN, which is designed to generate code for understanding digital twins.
- Explore performance improvements by integrating and benchmarking open-source AI models against proprietary alternatives.
- Develop energy-efficient multi-agent communication protocols to minimize unnecessary interactions and reduce overall computational cost.

EXPECTED RESULTS

- An improved version of the multi-agent AI system, LUMEN, enhanced using open-source and open-weight models.
- Validation will be conducted across use cases in the energy sector, utilizing a digital twin composed of data from 70 different smart meters, open-access hyperspectral imagery, 3D point cloud data from overhead power lines, and time series data from manufacturing processes.
- The results will be documented in a peer-reviewed scientific journal article.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[SINTEF](#) / Sagar Sen

Research Pillar: Green AI

G-AI.4 | How Green is Physics-Informed Machine Learning?

Keywords: physics-informed machine learning, physics-informed neural networks, computational efficiency, green AI, energy systems

STATE-OF-THE-ART

Physics-informed machine learning utilizes insights derived from (physical) theories during the development and training of machine learning models. Its potential to reduce the required amount of training data/duration and model complexity has a direct influence on the resulting model’s environmental footprint. Physics-informed machine learning (such as physics-informed neural networks, physics-guided neural networks, but also feature engineering and the inclusion of architecture constraints) thus not only improves generalization and extrapolation properties but is claimed to contribute to making AI greener.

SCIENTIFIC CHALLENGES

The scientific challenge addressed here is that the net benefit of physics-informed ML on energy efficiency has not been conclusively quantified yet, in particular for problems that require large simulation times. More generally, the trade-offs between energy consumption, model complexity, and accuracy achieved with physics-informed ML are not always obvious and rarely quantified.

RESEARCH ACTIVITIES

Physics-Informed Machine Learning (PIML) approaches show significant potential for enhancing the performance of learning models, particularly when the application can be partially or fully represented by a physical model. In this research, PIML shall be applied to solve a concrete learning problem from the energy sector (e.g., derive dynamic equivalents for power grids, dynamic simulation). Rather than simply solving the corresponding engineering problem, this research focuses on investigating if and how PIML can reduce the environmental footprint of machine learning. The research activities thus include measuring energy consumption of PIML models and how it is affected by the amount of included prior information and target accuracy

EXPECTED RESULTS

By collaborating with researchers from KNOW and INESC TEC, the project seeks to develop ML models that minimize environmental impact. Expected outcomes include i) a description of the methodology adopted for the integration of the physical model in the learning stage, ii) a characterization of the interplay between training data amount, model accuracy, and energy consumption, and iii) statements how the obtained results could generalize beyond the considered use case from the energy sector. Results of the project shall be published at an international academic venue.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[Know Center Research GmbH](#) / Dr. Franz Rohrhofer, Dr. Bernhard Geiger
Note: This challenge is co-hosted by [INESC TEC](#) / Dr. Francisco Fernandes

Research Pillar: Adaptive AI

A-AI.1 | AI-Enhanced Visual Understanding for Inspection and Maintenance

Keywords: Quality Inspection, Non-Intrusive Inspection, computer-vision, maintenance, Incremental learning, adaptive IA, Edge-computing

STATE-OF-THE-ART

As Industry 4.0 gains traction, enhancing quality inspection and defect detection has become essential across industrial sectors. AI and Deep learning have significantly advanced computer vision by improving tasks such as classification and object detection—making machine vision a key tool in industrial defect detection. However, and **despite the potential of computer vision to improve industrial operations, practical implementation remains challenging**. Many architectures developed in academic settings struggle to meet industrial demands due to high computational costs, energy consumption, and hardware requirements (e.g., CPU/GPU usage). In particular, **Non-intrusive inspection (NII)** for quality assessment is a vital component of intelligent manufacturing and machine vision-based surface defect detection should play a greater role in maintaining product quality by tackling the specific and nonstandard nature of industrial image data.

For this challenge we are considering various industrial assets - from pipelines and bridges to aircraft fuselages - that require frequent integrity checks for defects such as corrosion, cracks, delamination, or surface fatigue. Traditional inspection often relies on contact-based sensors or manual visual surveys that are costly, slow, and sometimes hazardous. **Non-intrusive inspection (NII)** powered by AI and computer vision offers a safer alternative, but today's models are usually trained for a single defect type, tightly coupled to one sensor setup, and brittle to real-world variability (lighting, material, viewpoint, surface coatings).

To unlock broad deployment, NII systems must become **adaptive**: able to generalize across asset types, incrementally learn new defect classes, and operate on resource-constrained edge devices where bandwidth, latency, or privacy limit cloud connectivity.

SCIENTIFIC CHALLENGES

The central objective of this challenge is to develop advanced **computer vision models for non-intrusive inspection (NII)**, capable of detecting surface or structural defects such as cracks, corrosion, or deformation in diverse real-world conditions. To achieve this, applicants may address a range of scientific challenges, including but not limited to:

1. **Robust Visual Modeling Across Domains**
Designing vision models that generalize across inspection targets (e.g., materials, geometries, lighting conditions) with minimal task-specific tuning.
2. **Incremental Learning for New Defect Types**
Enabling models to learn new defect categories or inspection tasks over time without retraining from scratch or forgetting prior knowledge.
3. **Learning from Limited or Imperfect Data**
Developing methods that perform well with sparse annotations, noisy labels, or rare defect instances—through self-supervised learning, weak supervision, or active sampling.
4. **Multi-Modal Fusion (Optional Enhancement)**
Exploring the integration of vision with other sensing modalities (e.g., thermal, ultrasonic, lidar) to improve detection sensitivity while maintaining non-intrusiveness.
5. **Model Efficiency for Real-World Deployment (Optional Enhancement)**
Ensuring models are computationally efficient and suitable for deployment on edge or embedded devices common in inspection scenarios (e.g., drones, robotic crawlers).
6. **Synthetic Data and Simulation-Aware Training**
Leveraging synthetic defect generation or simulated inspection environments to reduce reliance on large-scale annotated datasets.
7. **Confidence-Aware and Human-Assistive AI**
Incorporating uncertainty estimation or explainability features to support operator trust and guide human intervention when necessary.

These challenges are not exhaustive, and proposals exploring adjacent or complementary directions within the broader scope of AI-driven non-intrusive inspection are equally encouraged.

RESEARCH ACTIVITIES

Projects funded under this challenge are expected to follow a structured research plan focused on developing computer vision models for non-intrusive inspection. The research activities may include, but are not limited to, the following:

- **Problem Scoping and Application Definition:** Identify and clearly define the inspection context (e.g., corrosion detection, crack identification, surface anomaly inspection), the operational constraints, and the desired deployment environment.
- **Dataset Selection or Creation:** Select relevant publicly available datasets or create custom datasets aligned with the chosen inspection task. Proposals should outline data preprocessing strategies and address any limitations in data quantity or quality.
- **Model Development and Training:** Design and train computer vision models tailored to the selected application. This may include exploring convolutional or transformer-based architectures, foundation model adaptation, or lightweight models for embedded deployment.
- **Performance Validation and Evaluation:** Rigorously evaluate the model's accuracy, robustness, and generalization across different conditions or environments. Where applicable, performance should be assessed under real-world constraints (e.g., edge hardware, variable lighting).
- **Adaptive and Continual Learning Aspects (Preferred):** Introduce mechanisms that allow the model to incrementally learn new defect types or adapt to novel domains without retraining from scratch. This may include continual learning, self-supervision, or model update strategies.

These activities results are not exhaustive, and proposals exploring adjacent or complementary directions within the broader scope of AI-driven non-intrusive inspection are equally encouraged.

EXPECTED RESULTS

Projects are expected to deliver a fully validated computer-vision model for non-intrusive inspection that accurately detects the targeted defect class (e.g., corrosion, cracks) across varied real-world conditions. Outcomes should include: (i) a well-documented dataset or clearly referenced public dataset with accompanying preprocessing pipeline; (ii) a trained model that meets or exceeds baseline accuracy and recall while operating within the stated hardware- or latency constraints; (iii) an evaluation report demonstrating robustness to domain shifts (lighting, material, viewpoint) and, where proposed, successful incremental learning of at least one additional defect type with negligible performance loss on earlier tasks; and (iv) reproducible artifacts, code, trained weights, and inference scripts, released under an open or permissive license to facilitate further research and industrial uptake.

These expected results are not exhaustive, and proposals exploring adjacent or complementary directions within the broader scope of AI-driven non-intrusive inspection are equally encouraged.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[IMT](#) / Andon Tchechmedjiev

Note: this challenge is co-hosted by [PREDICT](#) / Leïla Belmerhnia

Research Pillar: Adaptive AI

A-AI.2 | Efficient Continual Learning Systems

Keywords: continual learning, foundation models, Parameter Efficient FineTuning (PEFT), Multimodality, Embedded AI

STATE-OF-THE-ART

Continual Learning equips AI systems to absorb new tasks and shifting data distributions while preserving prior competencies. At the same time, foundation models—notably Large Language Models (LLMs) and Vision-Language Models (VLMs)—have transformed the field by scaling across modalities and achieving strong zero-/few-shot generalization through massive pre-training. Yet these gains come with two pressing limitations: (1) the significant compute, memory, and energy budgets required to fine-tune or serve such models, and (2) pronounced catastrophic forgetting when they are naïvely adapted to streams of sequential tasks. These challenges are even more acute when models must operate on resource-constrained edge devices, where latency, power, and privacy concerns limit both retraining cycles and external data movement. Consequently, advancing continual-learning techniques that are both memory-efficient and hardware-aware offers a critical path toward making foundation models truly adaptive, whether in the cloud or at the edge, without sacrificing prior knowledge or prohibitive compute.

SCIENTIFIC CHALLENGES

The scientific challenges may include, but are not limited to, the following:

- **Continual Adaptation of Foundation Models:** Enabling parameter-efficient, streaming adaptation of large-scale pre-trained models (e.g., LLMs, VLMs) without retraining or full fine-tuning.
- **Resource-Aware Continual Learning:** Designing sparse and efficient CL algorithms that operate under strict compute, memory, and power constraints, especially for edge or embedded deployment.
- **Multi-Modal and Cross-Task Knowledge Integration:** Supporting lifelong learning across heterogeneous modalities
- **Adaptive Replay and Compressed Memory:** Developing principled buffer compression, prioritization, and retrieval strategies that scale to long horizon learning with bounded memory.
- **Evaluation Beyond Static Benchmarks:** Creating realistic, dynamic benchmarks that reflect open-world conditions, including non-stationary tasks, delayed supervision, and human interaction.

RESEARCH ACTIVITIES

The research activities may include, but are not limited to, the following:

- **Leverage Foundation Models in Continual Learning:** Investigate how large pre-trained models (e.g., LLMs, VLMs) can be adapted incrementally without retraining from scratch.
- **Integrate Parameter-Efficient Fine-Tuning (PEFT) Techniques:** Explore the use of adapters, LoRA, and other PEFT methods within continual learning pipelines to enable scalable and efficient adaptation.
- **Pursue Efficiency Through Architectural and Algorithmic Techniques:** Examine the role of sparsity, pruning, prompting, and modular designs in reducing the computational and memory overhead of continual learning.
- **Assess the Role of Multi-Modality:** Analyse how combining or separating modalities (e.g., vision and language) impacts knowledge transfer, forgetting, and adaptation across tasks.
- **Develop and Validate Novel Methodologies:** Propose new continual learning strategies informed by empirical evaluation, aiming to outperform current state-of-the-art approaches.
- **Incorporate Hardware-Aware Optimization:** Design continual learning methods with awareness of hardware constraints such as memory bandwidth, power consumption, and inference latency, ensuring solutions are viable for deployment on edge devices or resource-limited platforms.

EXPECTED RESULTS

Projects are expected to deliver novel continual learning methodologies that are scalable, efficient, and robust in real-world, dynamic environments. Solutions should demonstrate improved performance over current state-of-the-art techniques, particularly in terms of adaptability, memory efficiency, and reduced forgetting. Where applicable, methods should leverage foundation models and incorporate parameter-efficient updates suitable for sequential learning. Proposals are also encouraged to address hardware-awareness by quantifying computational cost, latency, and memory usage to ensure feasibility on edge or constrained platforms. Outcomes may include open-source codebases, benchmark datasets, evaluation tools, and peer-reviewed publications that advance the continual learning community.

These expected results are not exhaustive, and proposals exploring adjacent or complementary directions within this broader scope are equally encouraged

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

IMT
[TU/e/](#)

Research Pillar: Adaptive AI

A-AI.3 | Lifelong Learning and Fine-tuning in LLMs

Keywords: continual learning, foundation models, LLMs, Multi-modal LLMs, Reinforcement Learning

STATE-OF-THE-ART

Large Language Models (LLMs) and Multimodal Large Language Model (MLLM) offer strong zero-/few-shot abilities after massive pre-training. Yet each time they meet a new task or data stream, naïve fine-tuning is slow, costly, and prone to forgetting. Supervised fine-tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) align these models with human intent, but each adaptation pass is typically treated as a one-off, computationally expensive event. Bringing true lifelong adaptation to the foundation models—while keeping efficiency in check—is an important objective.

SCIENTIFIC CHALLENGES

The scientific challenges may include, but are not limited to, the following:

- Adaptation to dynamic environment – models for new domains or tasks without full retraining.
- Improve generalization and Mitigate forgetting - Improve performance on novel data while protecting previous knowledge.
- Multi-modality – The interplay between the different modalities such as vision, video and language in generalization
- Efficiency – Achieve model updates without increasing computation resources

RESEARCH ACTIVITIES

The research activities may include, but are not limited to, the following:

- Exploring continual pretraining and fine-tuning capabilities in LLMs and Multi-modal foundation models including modalities such as vision, video
- Designing novel continual and fine-tuning techniques to adapt to new data and domains
- Exploring continual fine-tuning in Reinforcement Learning from Human Feedback (RLHF) setting
- Comparing against standard benchmarks and metrics that capture forgetting and adaptation speed under real-world data drift.

EXPECTED RESULTS

Projects are expected to deliver:

- Analysis and evaluation on generalization and continual learning in LLMs and multi-modal LLM
- A novel design or methodology for effective and efficient learning in LLMs
- Public benchmarks, code, that move the community toward truly adaptive, efficient language models.
- Publications in top-tier A/A* conferences or journals.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

TU/e, [Data & AI Cluster](#), Department of Mathematics and Computer Science / Shruthi Gowda

Research Pillar: Adaptive AI

A-AI.4 | Flexible plastics sorting system based on the combination of multiple vision sensors and adaptive AI

Keywords: Automatic plastics sorting, vision sensors, adaptive AI

STATE-OF-THE-ART

Huge volumes of plastic are being generated every year, leading to worldwide environmental and health hazards throughout their lifecycle, from production to disposal. To tackle these challenges, significant efforts are being made to sort plastics waste to effectively recycle it. Even if these sorting operations are generally manually performed, automated sorting technologies are being developed to increase their efficiency and robustness [1]. Current developments show that sensor-based solutions combined with AI are promising and forthcoming [2, 3]. One interesting family of approaches relies on conformal prediction (CP). CP is a probabilistic framework for uncertainty quantification agnostic to any specific classifier that operate on the a posteriori probability distribution. CP allows to set a particular level of confidence required for a decision to be made, thus offering a way of avoiding misclassifications due to excessive uncertainty about the decision. CP can also be exploited for semi-supervised learning by selecting unlabelled datapoints for which high confidence predictions can be derived [REF]. CP has shown promise in for plastics sorting, particularly in the context of high-speed sorting, where incorporating uncertainty is paramount, although the technique has only been used with classical machine learning approaches.

References

- [1] Lubongo, Cesar, Mohammed AA Bin Daej, and Paschalis Alexandridis. "Automated sorting technology for plastic waste." *Reuse of Plastic Waste in Eco-Efficient Concrete*. Woodhead Publishing, 2024. 13-35.
- [2] Monica Moroni, Marco Balsi, Soufyane Bouchelaghem, Plastics detection and sorting using hyperspectral sensing and machine learning algorithms, *Waste Management*, Volume 203, 2025.
- [3] G. Maier, R. Gruna, T. Längle and J. Beyerer, "A Survey of the State of the Art in Sensor-Based Sorting Technology and Research," in *IEEE Access*, vol. 12, pp. 6473-6493, 2024

SCIENTIFIC CHALLENGES

Automated plastic waste sorting faces persistent challenges due to the high variability of plastic materials in the recycling stream. Transparent, black, or multi-polymer plastics are particularly difficult to identify using standard vision systems. While hyperspectral and infrared sensing, combined with AI, have improved material discrimination, building robust models that generalize across diverse, contaminated, or novel waste remains a significant hurdle.

Another major limitation is the inability of conventional models to adapt to new or unseen plastic types without full retraining. This restricts their long-term scalability in real-world settings where plastic packaging designs and materials evolve constantly. Vision foundation models (e.g., CLIP) offer a promising alternative by enabling zero- or few-shot classification via natural language prompts. However, these models lack built-in mechanisms to quantify uncertainty, which is essential for safety and reliability in high-throughput environments.

Conformal Prediction (CP) provides a model-agnostic framework for uncertainty quantification by generating prediction sets that meet predefined confidence levels. In the context of plastics sorting, CP can flag ambiguous items for manual review rather than risking misclassification. It also supports semi-supervised learning by identifying high-confidence predictions among unlabeled samples, enabling continuous system improvement without exhaustive annotation.

Integrating CP into foundation models introduces new research challenges. These models often operate under domain shift relative to their training data, potentially violating CP's assumptions and degrading its guarantees. Novel calibration strategies, such as domain-adaptive or transductive CP, must be explored to ensure validity. Additionally, defining appropriate nonconformity scores in multimodal contexts (e.g., combining image and prompt-based outputs) is an open area of investigation.

Finally, long-term deployment demands systems that can adapt continually while maintaining reliability. Combining CP with continual learning frameworks could allow plastic sorters to refine their predictions over time and detect distributional drift. However, ensuring that confidence estimates remain valid across evolving model states and calibration sets remains a largely unsolved problem, calling for new theoretical and algorithmic advances at the intersection of uncertainty quantification and adaptive learning.

RESEARCH ACTIVITIES

- **Review of Methods and Literature:** Conduct a focused review of recent developments in conformal prediction, particularly its use in computer vision, semi-supervised learning, and integration with foundation models.
- **Development of Baseline Sorting Models:** Set up baseline plastic sorting pipelines using existing vision models and sensor data (e.g., infrared or hyperspectral), to establish performance benchmarks.
- **Integration of Conformal Prediction:** Apply conformal prediction techniques to selected models to enable uncertainty-aware decision-making, including selective classification and confidence-controlled deferral.
- **Exploration of Zero-/Few-Shot Learning Approaches:** Investigate the use of vision foundation models (e.g., CLIP) for identifying novel plastic types, and evaluate how conformal prediction can regulate uncertainty in these settings.
- **Assessment of Semi-Supervised and Adaptive Learning:** Explore the potential of conformal prediction to support semi-supervised learning and continual model adaptation in evolving plastic waste environments.
- **Evaluation and Knowledge Transfer:** Analyze results in terms of reliability, adaptability, and sorting accuracy, and prepare deliverables including a report, code repository, and potential research outputs.

EXPECTED RESULTS

A novel adaptive AI method should be developed to detect and classify a wide range of plastic waste types using data from vision sensors operating across multiple wavelengths, including the near-infrared range (900–1700 nm). The method should be capable of adapting to new or evolving waste types not seen during initial training. Integration of conformal prediction is encouraged to enable uncertainty-aware decision-making, selective classification, and semi-supervised refinement. Applicants may also explore the use of vision foundation models to support zero-shot or few-shot learning for emerging material categories.

While limited real-world data may be provided, applicants are encouraged to supplement this using publicly available datasets, such as the Tecnalia WEEE Hyperspectral Dataset, or generate additional data as needed. The host institution can also provide access to conveyor belt setups and hyperspectral sensors for experimental validation of the proposed approach, which may require on-site testing.

The research is expected to result in at least one peer-reviewed scientific publication. Public release of code, models, and datasets developed during the project is strongly encouraged to ensure reproducibility and maximize impact.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[DTI](#) - Danish Technological Institute / François Picard

[IMT](#)

[TU/e](#) / Shruthi Gowda

Research Pillar: Adaptive AI

A-AI-IDM.5 | Batch pattern mining in industrial signals

Keywords: Time series, match & label, condition-based monitoring

STATE-OF-THE-ART

Industry 4.0 technologies enable easier access to operational data by leveraging various digital solutions that record key performance indicators and machine parameters. As a result, industrial operators are increasingly adopting condition-based maintenance (CBM) strategies to gain a better understanding of equipment wear and failure mechanisms. This concept, referred to as Prognostics and Health Management (PHM), has been identified as a critical component for optimizing unit downtimes and enhancing facility safety. A key and recurring task in PHM implementation is the contextualization phase, which involves labelling machine operating states by extracting business indicators from raw sensors measurements. From a machine learning standpoint, this task corresponds to pattern identification and classification challenges involving time series data of indefinite length and high sampling rates. However, the application of classical supervised AI techniques remains difficult in industrial contexts due to the high variability of these processes and the lack of labelled data. Unsupervised approaches have been extensively explored but fail to accurately capture variability in the timing and duration of the patterns, due to human involvement that leads to conditions that are hard to predict. Moreover, classical methods based on Dynamic Time Warping are slow and are difficult to scale over very long sequences. In this challenge, we want to explore the possibilities offered by state-of-the-art deep learning architectures for this problem, exploring either transfer learning through adaptation of foundational models for various types of signals, or synthetic data generations approaches whereby a large dataset of deformed reference patterns is generated and in turn serves to train a deep learning model. For example we could seek to explore generative architectures for signal forecasting (e.g. Crossformer).

SCIENTIFIC CHALLENGES

- Operationalize or design a signal foundational model
- Develop a self-supervised training paradigm to fine-tune the model to recognize or forecast reference patterns and sub patterns, in a situation where only one “ideal” example exists
- Adapt the model to an online setting to apply it on real industrial data without a labelled reference (then evaluated by experts). Show replication on two different industrial datasets provided by PREDICT.

RESEARCH ACTIVITIES

- Analyse the state of the art
- Develop a data augmentation strategy and associated self-supervision protocol
- Develop a forecasting deep learning model
- Evaluate on real industrial data

EXPECTED RESULTS

- 1 scientific publication (conference, workshops).
- 1 extended publication on top level journal in related domains.
- An IA-based algorithm that improves the performance of the contextualisation task of time series from an industrial process.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[IMT](#) / Andon Tchechmedjiev

Note: this challenge is co-hosted by [PREDICT](#) / Leïla Belmerhnia

Research Pillar: Human-Centric AI

HC-AI.1 | Empowering Human-AI Decision-Making through Real-Time User Modeling

Keywords: Human-AI interaction, AI-assisted decision-making, User modeling, Cognitive states, Behavioral experiments

STATE-OF-THE-ART

Artificial intelligence (AI) has been increasingly introduced to decision-making tasks that are traditionally done by humans, from selecting job candidates to diagnosing diseases based on clinical images. While fully automated decisions by AI are desirable for some applications, for more critical domains (e.g., healthcare, security, energy), it is more common to leave the final decisions to human experts who receive advice from AI systems. When these “AI-assisted decision-making” scenarios work well, the human-AI hybrid system can utilize the speed and efficiency advantages of AI, but at the same time allow human oversight for ethical and accountability considerations [1]. In tasks where both humans and AI algorithms are not impeccable, it is believed that human-AI shared decision-making can lead to the so-called “human-AI complementarity”, i.e., the combined performance is better than human or AI alone [2].

However, in order to reach these potentials, human experts need to trust the AI systems at an appropriate level and to rely on them at the right moments. Unfortunately, both over-reliance and under-reliance issues are frequently reported and achieving complementarity has been shown to be challenging [3]. Trust calibration is especially difficult when people work with AI systems, which are often by their design non-transparent in their processes and stochastic with regard to their outputs. To address these challenges, the emerging field of human-AI decision-making has made progress in at least four strands. First, there is a profusion of “explainable AI (XAI)” methods that allow “black box” models to explain their behaviors to the human users, with the hope that explanations foster trust [4]. Second, different uncertainty representation approaches have been used to help users critically evaluate AI suggestions on a case-by-case basis [5]. Third, human-computer interaction (HCI) researchers have highlighted the importance of users’ mental models of AI and invented design guidelines and patterns for building accurate mental models [6]. Finally, there is a sturdy progress in standardizing measurements for trust and reliance, both through self-report questionnaires and behavioral metrics during the decision-making tasks [7].

Most of the research efforts above have focused on one side of the human-AI collaboration problem, i.e., how to help users better understand AI, but overlooked the other side, i.e., how we can develop AI systems that understand more about their human users [8]. The promise of the latter approach is that by making AI systems more aware of users’ personal traits and mental states (e.g., decision-making style, confidence, cognitive load, etc.), the systems can adapt their inputs to the users for making decisions. The proposed challenge focuses on the potential of real-time user modeling for empowering human-AI decision-making.

References:

- [1] Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023, June). Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 1369-1385).
- [2] Rastogi, C., Leqi, L., Holstein, K., & Heidari, H. (2023, November). A taxonomy of human and ML strengths in decision-making to investigate human-ML complementarity. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 11, pp. 127-139).
- [3] Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 1-11.
- [4] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- [5] Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., ... & Xiang, A. (2021, July). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 401-413).
- [6] Google PAIR. 2019. People + AI Guidebook. pair.withgoogle.com/guidebook
- [7] Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., & Seaborn, K. (2022, April). Trust in human-AI interaction: Scoping out models, measures, and methods. In *CHI conference on human factors in computing systems extended abstracts* (pp. 1-7).
- [8] Steyvers, M., & Kumar, A. (2024). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, 19(5), 722-734.

SCIENTIFIC CHALLENGES

The applicant may choose to address one or more of the following challenges:

- How can AI understand users’ cognitive states, which are mostly latent variables that cannot be observed directly?
- How to model users when there is limited data available regarding a specific individual user?
- When we have a user model, can it be implemented in an AI system to improve human-AI shared decision-making performance? How does it influence user trust and user experience?
- How to design and execute user studies with actual domain experts who are difficult to find (problem of small sample size)?
- How to measure trust and reliance in user experiments that can be generalized to real-world applications?

RESEARCH ACTIVITIES

Research activities may include but not limited to one or more of the following:

- New methods for cognitive computational modeling of AI users in decision-making tasks
- New ways of collecting process (e.g., eye-tracking) or physiological data for user modeling
- A combination of theory-based cognitive modeling and data-driven user modeling
- Empirical studies to identify important cognitive variables in human-AI decision-making
- New methods for measuring trust and reliance on AI among human experts
- Applied research on human experts in the context of real-world applications

EXPECTED RESULTS

The research by the exchange researcher will move the emerging field of user modeling in human-AI decision-making forward in one of the following areas, including new user models, new data collection methods, system evaluation results, or methodological contributions (e.g., a new trust measurement). We expect the research results to be published in a top conference in HCI (e.g., CHI, IUI, UMAP, etc.).

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[TU/e](#) / Dr. Chao Zhang

Research Pillar: Human-Centric AI

HC-AI.2 | Interpretability and uncertainty in predictive models

Keywords: Explainable AI, Uncertainty Quantification, Conformal Prediction, Trustworthy AI

STATE-OF-THE-ART

AI models that score high on both accuracy and interpretability are in demand. Traditional post-hoc explanation methods often suffer from inconsistencies and lack of fidelity to the underlying models, which has led to increased emphasis on developing models that are more transparent and interpretable by design. Simultaneously, accurate uncertainty quantification in predictive models is essential not only for informed decision-making, but also for instilling trust in automated systems and ensuring their safety [1]. Failure to properly account for uncertainty can lead decision-makers to inappropriately trust potentially incorrect outputs, undermining confidence even in accurate predictions.

While current approaches often treat interpretability and uncertainty quantification objectives in isolation, there is interesting recent work on bringing the two domains closer together. For instance, it is possible to use the framework of Conformal Prediction (CP) [3] – a robust, model-agnostic approach to uncertainty quantification – to quantify the uncertainty associated with explanations [4]. Besides such uncertainty-aware explanations, there have also been recent advances in the design of intrinsically interpretable models with guaranteed accuracy of explanations [5, 6], employing sparse representations to evade problems with first learning a complex model and then finding post-hoc explanations..

References:

- [1] Bhatt, Umang, et al. "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty." Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021.
- [2] Antoran, Javier, et al. "Getting a CLUE: A Method for Explaining Uncertainty Estimates." International Conference on Learning Representations. 2020.
- [3] Vovk, Vladimir, et al. "Algorithmic learning in a random world." Vol. 29. New York: Springer, 2005.
- [4] Marx, Charles, et al. "But Are You Sure? An Uncertainty-Aware Perspective on Explainable AI" Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, PMLR 206:7375-7391, 2023.
- [5] Swamy, Vinitra, et al. "The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations". *arXiv preprint arXiv:2307.00364* (2023).
- [6] Swamy, Vinitra, et al. "InterpretCC: Conditional Computation for Inherently Interpretable Neural Networks." *arXiv preprint arXiv:2402.02933* (2024).
- [7] Example datasets:
 - Barlacchi, Gianni, et al. "A multi-source dataset of urban life in the city of Milan and the Province of Trentino", Scientific Data, 2015, <https://doi.org/10.1038/sdata.2015.55>.
 - Anderson, Eric W., and Caleb Phillips, CROWDAD CU/ANTENNA, IEEE DataPort, 2009, <https://doi.org/10.15783/C7VC7V>.
 - Sun, Yanzan, et al. "An Open Source Wireless Communication Database for Radio Access Network", Industrial Networks and Intelligent Systems, 2023, https://doi.org/10.1007/978-3-031-47359-3_5.

SCIENTIFIC CHALLENGES

The applicant may choose to address one or more of the following challenges:

- Developing new models that integrate interpretability and uncertainty quantification, enabling human-centric, trustworthy predictions and explanations.
- Exploring and extending existing methodologies like InterpretCC for various prediction tasks with relevance for telecom applications, for instance multivariate, irregular timeseries forecasting.
- Creating metrics and evaluating real-world applications to assess how interpretability and/or uncertainty information influence users' trust and the usability of critical systems.

RESEARCH ACTIVITIES

- **Literature review:** Study existing methods combining interpretability and uncertainty quantification.
- **Method development and benchmarking:** Develop and evaluate methods suitable for various telecom-relevant prediction tasks, benchmarking them on synthetic and open-source datasets – for examples, see Ref. [7].
- **Real-world evaluation:** Testing selected interpretability and uncertainty quantification methods on users, measuring their impact on tasks like forecasting and anomaly detection.

EXPECTED RESULTS

We expect the exchange to provide valuable contributions to:

- Refined methods: Improving ML algorithms for prediction tasks, with integrated interpretability and uncertainty features.

- Practical applications: Demonstrating reliability in real-world use cases such as time series forecasting and anomaly detection.
- Human-centric trust: Enhancing user trust in AI systems through accurate uncertainty estimation and interpretable methodologies.
- Dissemination: Publishing findings in a high-impact conference or journal to engage the research community and drive advancements.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[TELENOR](#) / Jeriek Van den Abeele

Research Pillar: Human-Centric AI

HC-AI.3 | Human-Centric Explainable AI for Critical Digital Infrastructure

Keywords: Interpretability; Anomaly Detection; Intrusion Detection; Time-Series Forecasting; Uncertainty Quantification

STATE-OF-THE-ART

Operators across critical infrastructure sectors – telecommunications, energy grids, water treatment, and transportation networks – increasingly adopt AI for monitoring, security, and service reliability. However, opaque models hinder trustworthy adoption, and often solutions fall short of addressing practical, domain-specific requirements. In the telecom domain, semi-supervised neural networks have proven useful for O-RAN traffic anomaly detection, with the XAIomaly framework [Basaran et al., 2025] combining a contractive autoencoder with efficient Shapley value estimation to highlight the features driving each alarm. In intrusion detection, neural networks can be employed to spot attacks on IoT networks, with post-hoc explanation techniques like SHAP and LIME revealing why particular flows are flagged [Sharma et al., 2024]. Nevertheless, delivering real-time, high-throughput explanations for streaming network timeseries data, typically multivariate, remains challenging. For service reliability forecasting, concept-based time-series models could offer transparent “nearest-neighbour” explanations [Obermair et al., 2023], while conformal prediction methods could be used to attach well-calibrated uncertainty intervals [Stankeviciute et al., 2021], though it does not seem like these approaches have so far been used for telecom-specific applications. While conformal prediction methods could be used to attach well-calibrated uncertainty intervals [Stankeviciute et al., 2021], though these approaches have so far not gained traction for telecom applications.

SCIENTIFIC CHALLENGES

We invite proposals addressing one or more of the issues introduced above, for example:

- **Explainable anomaly detection:** Developing explainable anomaly detection models that pair high detection accuracy on telecom KPI streams with relevant explanation metrics
- **XAI benchmarking for cybersecurity:** Automating evaluation of post-hoc explainers (SHAP, LIME, Integrated Gradients, ...) on unsupervised intrusion detection models, and assessing faithfulness, stability, and completeness [Nauta et al., 2022] of the applied methods [Al et al., 2025]
- **Interpretable forecasting:** Designing concept-based forecasting models that produce human-intelligible explanations, e.g., by model-agnostic prototypes [Obermair et al., 2023]
- **Forecasting with uncertainty quantification:** Providing rigorously calibrated prediction intervals for critical service metrics

RESEARCH ACTIVITIES

Relevant research activities could include:

- **Landscape mapping:** Conducting a targeted literature review of existing approaches, highlighting gaps and relevant opportunities
- **Method design:** Developing one or more approaches, e.g., autoencoder+SHAP benchmarking suite or forecaster+conformal intervals, specifying algorithms and evaluation metrics
- **Prototype implementation:** Building modular, open-source tooling for detection+explanation, XAI benchmarking, or forecasting+uncertainty
- **Quantitative evaluation:** Injecting synthetic anomalies or attacks; measuring detection/false-alarm rates; explanation faithfulness/stability; forecast accuracy; uncertainty calibration, using open telecom time-series

and network-security datasets (e.g., Numenta Anomaly Benchmark, Telecom Italia Open Data, NSL-KDD, CIC-IDS2017).

EXPECTED RESULTS

Valuable contributions from the exchange can for example include:

- Novel or refined methods: Devising or improving ML algorithms for anomaly or intrusion detection, with integrated explainability or uncertainty features.
- Practical applications: Demonstrating reliability of the proposed methods in domain-specific use cases based on real-world data
- Open-access artifacts: Public code, datasets, and reproducible benchmarks to enable community adoption
- Dissemination materials and activities

At least one peer-reviewed publication at an international academic venue or in a high-impact journal is expected from the exchange.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[TELENOR](#) / Jeriek Van den Abeele

Research Pillar: Human-Centric AI

HC-AI.4 | Human-Centered Causal Machine Learning for Real-World Decision Support

Keywords: Causal Machine Learning, Explainable AI (XAI), Counterfactual reasoning, Image analysis

STATE-OF-THE-ART

Recent advancements in machine learning (ML) have demonstrated significant impact across various domains by enabling predictive modelling, pattern recognition, and automation at scale. The rise of Explainable AI (XAI) and Causal Machine Learning (CML) [1,2] is shifting the focus from pure correlation-based learning to models that can reason about cause-effect relationships. Causal Machine Learning aims to support actionable, human-centered decisions by enabling models to answer questions like “What if we change this intervention?” or “What caused this outcome?”, rather than just predicting outcomes. This is especially relevant for decision-critical domains where understanding why an outcome occurs is as important as predicting what will happen.

Traditional ML approaches in fields such as remote sensing or healthcare predominantly rely on supervised or unsupervised learning models that capture statistical correlations. However, these models often fall short in guiding interventions or supporting policy-level decisions, which require a deeper understanding of causal mechanisms. Causal Inference methods such as Structural Causal Models (SCM) and Double Machine Learning (DML) have matured within economics and business intelligence, but their integration into broader ML pipelines remains nascent. Moreover, the combination of CML with XAI methods, in particular counterfactual reasoning [3,4] is still emerging and presents an exciting frontier for building explainable, decision-focused AI systems.

References

- [1] Schölkopf, B., et al. (2021). Towards Causal Representation Learning. *Proceedings of the IEEE*, 109(5), 612–634.
 [2] Kaddour, J. et.al. Causal Machine Learning: A Survey and Open Problems. *arXiv:2206.15475*
 [3] Pawlowski, N., et al. (2020). Deep Structural Causal Models for Tractable Counterfactual Inference. *Advances in Neural Information Processing Systems (NeurIPS)*.
 [4] Janzing, D., Minorics, L., & Blöbaum, P. (2019). Feature relevance quantification in explainable AI: A causal problem. *arXiv:1910.13413*

SCIENTIFIC CHALLENGES

Despite the theoretical progress in causality, several scientific and technical challenges hinder the practical deployment of Causal Machine Learning and XAI:

- Integration with ML workflows: Bridging traditional statistical causal methods with modern ML techniques remains a key challenge.
- Counterfactual Reasoning over Images: Generating counterfactuals for visual inputs, i.e., what an image would look like under a different treatment, requires hybrid models combining vision and causality.
- Personalization of causal inference: Adapting causal conclusions to individual users or environments (e.g., farmers, patients, consumers) demands the development of personalized treatment effect estimation models.
- Scalability and generalizability: CML models must be scalable to big data (e.g., Earth Observation or sensor data) while maintaining the ability to generalize across time and spatial scales.

RESEARCH ACTIVITIES

The research activities will focus on the investigation of Causal Machine Learning models enhanced with image analysis to enable personalized, interpretable, and actionable decision support. A key application will be to develop a model to explore the causal effects of farming practices on yield productivity using Earth Observation (EO) imagery. Key research activities include:

- Development of a modular CML pipeline that integrates EO image analysis and domain knowledge with structural causal models capable of simulating counterfactual outcomes.

- Use of CML pipeline for individual fields, enabling farmers to compare observed yield outcomes with hypothetical alternatives (e.g., different irrigation strategies), supporting personalized recommendations and adaptive planning.
- Comparison with baseline predictive ML models, demonstrating the added value of causal and counterfactual reasoning in generating insights that are both interpretable and actionable.
- Visual interpretation of causal effects, enabling localized, human-understandable recommendations.

EXPECTED RESULTS

Advance the research in the field of human-centered AI in the proposed topic, in particular:

- Demonstrate the feasibility and benefits of integrating CML into image analysis workflows.
- Provide personalized, explainable insights to end-users, supporting human-centered AI in critical sectors such as agriculture.
- Enable data-driven intervention strategies, increasing efficiency and productivity in industries reliant on large-scale observational data.

The research results have the potential to be applied to several domains, such as agriculture, health or manufacturing. In the case of agriculture that the topic focuses on, the results will support farmers by identifying the most effective practices under varying conditions, leading to improved yield, optimized resource use, and increased economic returns.

We expect the research results to be published in a top conference or journal.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[National University of Science and Technology POLITEHNICA Bucurest \(UPB\)](#) / Prof. Adina Magda Florea

Research Pillar: Human-Centric AI

HC-AI.5 | Explainable AI Methods for Human-AI Decision Making

Keywords: Explainable AI (XAI), Spatio-temporal models, Video analysis

STATE-OF-THE-ART

Recent advances in deep learning, particularly multi-modal Transformers, have shown strong potential for modeling complex spatio-temporal dependencies in tasks such as Human Action Recognition (HAR). This is especially relevant in real-world environments like healthcare, ambient assisted living (AAL), and smart manufacturing, where accurate and interpretable behavior analysis is critical.

A variety of deep neural network architectures have been explored for HAR, including Temporal Convolutional Networks (TCNs), Spatial-Temporal Graph Convolutional Networks (ST-GCNs), and Transformer-based models that explicitly capture both spatial and temporal attention patterns. Notable examples include the Video Action Transformer Network [1], ConvTransformer, and Spatio-Temporal Attention Networks, which have achieved state-of-the-art performance on benchmark datasets.

However, these models often suffer from a lack of interpretability, limiting their adoption in domains where trustworthy and transparent decision-making is essential. The intricate interaction between spatial and temporal dynamics introduces additional layers of complexity, making it difficult to understand how input features propagate through the network to influence the final prediction.

Despite growing interest in explainable AI (XAI), there are still relatively few methods tailored to explain why these models succeed or fail, particularly in HAR from video data [2]. Recent trends point toward counterfactual reasoning [3], causal attribution, and attention-based explanation mechanisms [4, 5] as promising directions for addressing these challenges, enabling more robust, interpretable, and human-centered HAR systems.

References

- [1] Girdhar, R., et al. (2019). Video Action Transformer Network. CVPR Open access.
- [2] Chou Y.L. et al. Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications. arXiv:2103.04244
- [3] Li et al. (2025) – Transformer-Based Spatial Temporal Counterfactual Outcomes Estimation. arXiv:2506.21154
- [4] Abnar, S., and Zuidema, W. (2020). Quantifying Attention Flow in Transformers. In 58th Annual Meeting of the Association for Computational Linguistics
- [5] Wang and Liu (2024) – STAA: Spatio Temporal Attention Attribution for Real Time Video Models. arXiv:2411.00630

SCIENTIFIC CHALLENGES

Existing XAI methods often fall short in capturing the intertwined spatial and temporal dependencies, requiring novel or hybrid techniques that are model-aware yet user-friendly.

- Aligning explainability outputs with the expectations of end-users and identifying which types of explanations (e.g., counterfactual, attention-based) are most useful in different real-world decision-making contexts.
- New explanation methods that offer transparency without significantly increasing the computational or architectural complexity of already large Transformer models. Lightweight XAI modules or post-hoc causal probes are needed to maintain scalability in time-sensitive applications like HAR, autonomous driving, or video surveillance.
- Metrics to evaluating the relevance and utility of explanations in dynamic spatio-temporal settings.

RESEARCH ACTIVITIES

- Develop Explainable Algorithms for Multi-Modal Spatio-Temporal Transformers. Leverage internal attention dynamics and incorporate structured priors to uncover how spatial, temporal, and modality-specific inputs contribute to the model's decisions.
- Design XAI Methods that capture temporal and causal dynamics. Develop XAI techniques that go beyond static feature attribution by modelling the flow of information over time, such as generating temporally-aware saliency maps, attention trajectories, and counterfactual timelines.

- Evaluate explanation quality across human-centric dimensions. Establish a robust evaluation protocol combining quantitative metrics with qualitative assessments (e.g., user trust, usability)
- Explore the integration of counterfactual generation modules into existing HAR and video processing pipelines.

EXPECTED RESULTS

Advance the research in the field of human-centered AI in the proposed topic, in particular:

- Development of advanced explainability techniques for multi-modal spatio-temporal transformer models, with a focus on enhancing transparency and usability in real-world decision-making scenarios. These techniques will integrate attention-based explanations, causal attribution, and counterfactual reasoning to offer rich, interpretable insights into the behavior of complex deep learning systems.
- Contribution to human-in-the-loop systems where interpretability enhances model supervision and correction.
- Develop a use case for these models in one application domain, e.g. HAR.

We expect the research results to be published in a top conference or journal.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[National University of Science and Technology POLITEHNICA Bucurest \(UPB\)](#) / Prof. Adina Magda Florea

Research Pillar: Trustworthy AI

T-AI.1 | Security and Robustness of AI systems

Keywords: AI security/robustness, trustworthiness, LLMs, adversarial machine learning, verifiability, uncertainty quantification

STATE-OF-THE-ART

The rapid shift of the form of AI systems has introduced several challenges related to the security and robustness of such systems and consequently to their trustworthiness. The continuous growth and development of new techniques and methodologies in the AI field makes security of related systems even more difficult to achieve as the attack vectors are expanded with every new advancement in the field. Additionally, ensuring robust performance and accurate uncertainty quantification in AI models is crucial for maintaining trust in automated systems. The state of the art refers mainly to research on adversarial machine learning, relevant mitigation measures in the traditional machine learning field, and uncertainty estimation techniques like conformal prediction. It is of utmost importance to conduct research on the security and robustness of current state-of-the-art AI systems such as LLMs and indicate new possible attacks and/or provide relevant solutions.

SCIENTIFIC CHALLENGES

The main research challenges relate to: (1) adversarial machine learning attacks, (2) adversarial machine learning detection, (3) adversarial machine learning defences (4) LLMs related attacks such as prompt hacking or adversarial attacks, (5) LLMs defences, (6) AI fairness, (7) AI security by design approaches, (8) monitoring and measuring AI systems security (9) means for verifiable training and/or inference in AI systems, (10) uncertainty quantification in AI systems (e.g., timeseries forecasting) to improve robustness, and (11) calibration techniques to address biases and enhance model reliability.

RESEARCH ACTIVITIES

Research on any security or robustness aspect of traditional or modern AI systems. The activities expected will be related to all or some of the following: (1) identification of an interesting topic in the relevant research area, (2) literature survey for the selected topic, (3) proposal for a novel approach for that topic, (4) development of a proof of concept for the proposed approach, (5) comparison of the proposed approach with similar approaches in literature, (6) reasoning about the significance of the approach, (7) preparation of a relevant paper and (8) submission of the paper to a related venue.

EXPECTED RESULTS

The ENFIELD project will leverage novel scientific results to increase the trustworthiness of AI. By leveraging the results from this topic directions and guidelines towards the development of a trustworthy AI framework for EU will be facilitated. In addition to that, the involved partners and research will collaborate, exchange knowledge, and expertise to further develop their research activities and future collaborations. It is required to produce at least one scientific publication out of this collaboration.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[NTNU](#) / Georgios Spathoulas
[TELENOR](#) / Jeriek Van den Abeele

Research Pillar: Trustworthy AI

T-AI.2 | Privacy and Compliance of AI systems

Keywords: AI privacy, AI compliance, sensitive data, homomorphic encryption, federated learning

STATE-OF-THE-ART

AI systems process a vast amount of data (personal/sensitive). and they are strongly required to conform to the relevant regulations (GDPR). On top of that, it is important to advance state of the art with respect to technical measures that can facilitate privacy protection of data and AI models. The recent advancements of LLMs have brought new issues with regards to training data availability, consent of users to have their personal data included in training datasets and access control to closed access AI models the use of which is offered under an AI as a Service scheme.

SCIENTIFIC CHALLENGES

As AI systems become prominent in our lives it is important to deal with privacy requirements and enable regulations that can be practically applied to real world systems. The main research challenges relate to: (1) identification of privacy leakage in AI systems, (2) methodologies to make AI systems more privacy friendly, (3) use of cryptographic techniques (homomorphic encryption, zero knowledge proofs) to enhance privacy, (4) use of federated learning approaches to increase privacy in distributed setups, (5) regulatory framework for AI systems, (6) tooling to monitor/prove regulatory compliance, (7) users perspective on trustworthy AI and relevant privacy issues and (8) approaches to increase users' awareness.

RESEARCH ACTIVITIES

Research on any privacy and regulatory aspect of traditional or modern AI systems. The activities expected will be related to all or some of the following: (1) identification of an interesting topic in the relevant research area, (2) literature survey for the selected topic, (3) proposal for a novel approach for that topic, (4) development of a proof of concept for the proposed approach, (5) comparison of the proposed approach with similar approaches in literature, (6) reasoning about the significance of the approach, (7) preparation of a relevant paper and (8) submission of the paper to a related venue.

EXPECTED RESULTS

ENFIELD project expects that the research work towards the specific challenge will provide advancements to achieving privacy preservation (to the extent that this is feasible) in AI systems. Alternatively, collaboration can enhance regulatory frameworks that are coming up globally and provide tooling relevant to their practical application. The researcher may work towards novel AI workflows and approaches that will facilitate the development of trustworthy AI systems that are compliant with such frameworks. The project expects at least one scientific publication in one of the mentioned areas as the outcome of the exchange and the formation of a collaboration between the visiting researcher and the hosting institution that can be extended even after the research visit.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[NTNU](#) / Georgios Spathoulas

[TUC](#) - Distributed and Self-organizing Systems Group / Dr. Sebastian Heil

[TELENOR](#) / Jeriek Van den Abeele

Research Pillar: Trustworthy AI

T-AI.3 | AI in Distributed systems

Keywords: Trustworthy AI, AI Distributed Systems, Trust Modelling, Software Architecture, Method Engineering

STATE-OF-THE-ART

The integration of artificial intelligence is rapidly increasing and becoming more prevalent in our daily routines and the systems we interact with regularly, such as in healthcare, finance, transportation, social media, and online services. Those systems are utilizing AI to automate analysis and processing tasks and are becoming integral in decision-making processes. The underlying architecture of the systems is typically distributed, integrating AI services as system components or implementing a distributed AI system on its own. New trust-related challenges arise from the interplay of the distributed nature of the system architecture with the specific characteristics of AI components. These challenges need to be addressed in all phases of the system's lifecycle: in the architectural design of the system as well as during development, testing, maintenance and operation. Addressing the trust challenges at an early stage helps in creating a resilient architecture capable of supporting the dynamic and distributed nature of modern systems, ultimately enhancing the overall trustworthiness of the system. Existing modelling and analysis techniques for distributed systems lack a systematic consideration of trust aspects and AI-related characteristics.

SCIENTIFIC CHALLENGES

AI components are used in different parts of complex distributed or even federated systems, which raises challenges. One challenge that arises from integrating AI is that it impacts the trustworthiness of the overall system architecture. The unpredictability and opacity of AI components can harm users' trust in the system. This issue is particularly critical because AI models are often considered "black boxes", making it difficult to predict their behaviour and understand their decision-making processes. Building trustworthy AI systems requires implementing robust verification and validation techniques throughout AI training and inference phases. Ensuring the transparency, robustness, and fairness of AI systems is essential to maintaining user trust. Another challenge that arises due to integrating multiple AI components within a distributed system is that these components can interact with each other without clear verification or understanding, increasing the complexity of the entire system. These AI-to-AI interactions can be unpredictable and lack transparency, leading to potential issues in system reliability and trust. Increasing reliance of distributed systems on third-party AI services (AlaaS) poses another challenge. AI service consumers need to be able to ensure that the actual model and configuration used by the provider aligns with the intended model and configuration, which affects the system's verifiability and trustworthiness. Due to the nature of AI models, this cannot be verified on the results. Thus, the AlaaS providers can use an alternative model or manipulating the model parameters to reduce running costs and utilize less energy without the consumer's knowledge. Service Consumers currently have only limited means to verify the integrity and performance of third-party AI models. To address the above challenge of distributed AI systems, we are particularly interested but not limited to contributions in the following areas: Modelling and analysing trust on the architectural level; Users' perception of trust; Trust in Federated learning; Verifiability of AI inference in AlaaS scenarios; LLMs to assist trust and risk assessment; Explainable AI.

RESEARCH ACTIVITIES

We invite researchers to collaborate on one or more of the following research activities: the extension of a taxonomy of trust in distributed AI system architecture; the specification of a suitable visual modeling language; the development of infrastructure supporting the modeling; the design of algorithms for automatic analyses; the evaluation of trust modeling in distributed AI systems. Other research activities related to the challenges are described above.

EXPECTED RESULTS

We expect the exchange to provide valuable contributions to the long-term goal of designing a method for architectural trust modeling in complex distributed AI systems. The method will facilitate the creation of distributed “trustworthy by design” AI systems by enabling system architects to document and analyse trust in their architectural system blueprints. For researchers, the results will contribute to establishing a common vocabulary and representation of trust in distributed AI systems as a first step to consolidate the body of knowledge in this relatively young field and facilitate the communication and thus collaboration. The exchange also aims at fostering knowledge transfer and networking with other groups working in related fields such as information systems, distributed systems and software engineering.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[TUC](#) - Distributed and Self-organizing Systems Group / Dr. Sebastian Heil
[NTNU](#) / Georgios Spathoulas

Research Pillar: Trustworthy AI

T-AI.4 | Enhancing User Trust in AI-powered Chatbots using UBM

Keywords: User Trust, Chatbot Interaction, Trustworthy, User Behaviour

STATE-OF-THE-ART

AI-powered Chatbots are on the rise and have been integrated into various domains, including customer service, e-health, and digital assistants [1], [2]. However, user trust in these systems remains a significant challenge, particularly when chatbot behaviour feels rigid, slow, or disconnected from user expectations [3], [4]. Recent studies highlight that user trust is influenced by transparency, accuracy, emotional intelligence, and the chatbot's responsiveness to user behaviour [5]. Therefore, researchers began to highlight the importance of adaptive chatbots that adjust their behaviour and responses based on user behaviour to build user trust and enhance the user experience [5]. Currently, identifying users' behaviour typically involves monitoring their interactions with the chatbot as well as user logs to learn patterns, preferences, and intent over time. This often manual work requires time and effort to classify the user based on their behaviour and adapt the chatbot accordingly.

References

- [1] J. N. K. Wah, "Revolutionizing e-health: the transformative role of AI-powered hybrid chatbots in healthcare solutions," *Front. Public Health*, vol. 13, p. 1530799, Feb. 2025, doi: 10.3389/fpubh.2025.1530799.
- [2] T. P. Nagarhalli, V. Vaze, and N. K. Rana, "A Review of Current Trends in the Development of Chatbot Systems," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India: IEEE, Mar. 2020, pp. 706–710. doi: 10.1109/ICACCS48705.2020.9074420.
- [3] J. Li, L. Wu, J. Qi, Y. Zhang, Z. Wu, and S. Hu, "Determinants Affecting Consumer Trust in Communication With AI Chatbots: The Moderating Effect of Privacy Concerns," *J. Organ. End User Comput.*, vol. 35, no. 1, pp. 1–24, Aug. 2023, doi: 10.4018/JOEUC.328089.
- [4] S. Sousa, J. Cravino, and P. Martins, "Challenges and Trends in User Trust Discourse in AI," *Multimodal Technol. Interact.*, vol. 7, no. 2, p. 13, Jan. 2023, doi: 10.3390/mti7020013.
- [5] I. K. F. Haugeland, A. Følstad, C. Taylor, and C. A. Bjørkli, "Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design," *Int. J. Hum.-Comput. Stud.*, vol. 161, p. 102788, May 2022, doi: 10.1016/j.ijhcs.2022.102788.

SCIENTIFIC CHALLENGES

Building user trust in AI-powered chatbots remains a persistent challenge. One possible direction to address this issue is by personalising chatbot interactions based on real-time understanding of user interaction patterns. Researchers have begun to explore User Behaviour Modelling (UBM) to create adaptive interfaces and websites that dynamically change according to users' behaviour during their interactions. This method requires monitoring user logs during their sessions with the websites to detect user characteristics. However, those methods often require time to collect, analyse and classify the users. Utilising AI methods to model and classify users based on their interactions with the chatbot can speed up chatbot adaptation in real-time and enhance user trust and experience in a shorter timeframe. Another key challenge is that current UBMs often represent averaged user behaviours and fail to capture the unique individual characteristics of the user. Ignoring those personal characteristics and specific needs, making it essential to adapt UBMs to different user characteristics through low-level interaction assessment.

RESEARCH ACTIVITIES

We invite researchers to collaborate on one or more of the following research activities:

- Conducting an intensive literature survey to identify and analyse key characteristics of chatbots that influence user trust. This includes findings from existing studies to establish a framework of user trust aspects in AI-powered chatbots
- Conduct a comprehensive investigation into the characteristics of user behaviour and analyse their influence on user classification. This includes identifying behavioural patterns, extracting relevant features, and assessing how these behaviours contribute to or predict the user classification through statistical analysis and/or machine learning techniques
- Investigate and evaluate various methodologies for classifying and profiling users based on their behavioural data. This includes analysing machine learning and clustering approaches to discover different user classes and behavioural patterns

Develop and implement a low-level, real-time user behaviour modelling approach that captures individual user characteristics during early interactions

EXPECTED RESULTS

We expect the exchange to deliver the following results:

- An understanding of the user trust aspects and chatbot characteristics
- A novel framework to map chatbot characteristics to user behaviour and user trust aspects
- A novel algorithm to implement the approach and create a proof of concept
- One journal publication

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[TUC](#) - Distributed and Self-organizing Systems Group / Dr. Sebastian Heil

Research Pillar: Trustworthy AI

T-AI.5 | Trustworthy ML based scheduling for the energy domain

Keywords: Trustworthiness of AI; AI Robustness; Adversarial Machine Learning; Uncertainty Quantification; Energy Usage Monitoring.

STATE-OF-THE-ART

Production scheduling in modern manufacturing increasingly leverages AI for efficient resource allocation. However, trust in ML-driven scheduling is challenged by the fragility of models under adversarial conditions - subtle disruptions can cause inefficiencies without being easily detectable. Integrating such scheduling with a Digital Twin (DT) of a production line allows for simulation and controlled experimentation with AI robustness in a near-real environment. This research focuses on using this DT platform to explore trustworthiness of ML scheduling algorithms under realistic energy-aware constraints and adversarial perturbations.

SCIENTIFIC CHALLENGES

Develop an energy-aware ML-based scheduling system for a simplified digital twin of a production line (e.g., involving only 1-2 machines like *Linea Robot*, *Roland 307* and/or *Imbustatrice*). Simulate adversarial perturbations on selected scheduling inputs (e.g., task durations, order quantity or order priority) and observe their effects on scheduling outcomes. Design and implement basic robustness checks or recalibration mechanisms to mitigate the impact of these perturbations. Ensure energy consumption profiles are considered during optimization (e.g., penalise high energy schedules). Emphasis on one adversarial attack scenario and one mitigation/recalibration approach, not exhaustive coverage.

RESEARCH ACTIVITIES

The research will be undertaken by integrating a simplified digital twin model of a production line using data from one or two machines (e.g., *Linea Robot*, *Roland 307* and/or *Imbustatrice*) provided by Maggioli. Historical and real-time data on job orders, machine capacity, and energy consumption will be pre-processed to support scheduling. A machine learning-based production scheduler will be developed with energy-aware optimization objectives. An adversarial ML scenario will then be implemented by subtly perturbing selected input parameters such as job duration or order priority to simulate a realistic attack. The impact of these perturbations on scheduling performance (e.g., delays, energy usage, cost) will be evaluated. To address trustworthiness, a basic mitigation strategy such as a recalibration method or anomaly detection filter will be proposed and tested. The final phase will include documentation of results, preparation of scientific publication(s), and delivery of a lightweight toolkit for future use by the hosting institutions.

EXPECTED RESULTS

- A *proof-of-concept demonstration* of an adversarial attack on an ML-based scheduler in a digital twin setting.
- A simple but effective mitigation strategy to improve AI robustness.
- Metrics and insights on energy-aware scheduling under adversarial conditions.
- 1-2 scientific papers (e.g., targeting venues like *Trustworthy AI*, *Energy Informatics*, or *Industry 4.0 journals*).

Strengthened collaboration between researcher and Maggioli (industrial partner) & NTNU (academic partner) on adversarial testing in production systems.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[MAGGIOLI](#) / Astik Samal and Andrea Montefiori
[NTNU](#) / Georgios Spathoulas and Georgios Kavallieratos

Research Pillar: Trustworthy AI

T-AI.6 | Unified approach to assess safety and security of AI systems

Keywords: AI safety and security, Trustworthy AI, Risk assessment, Unified score.

STATE-OF-THE-ART

Artificial Intelligence (AI) is more and more present in our everyday life, demonstrating, as many disruptive technologies, a duality between the benefits of using it, and the potential dangers and actual harm it may lead to. Safety and security risks associated with AI systems must then be assessed to cope with AI incidents and hazards linked to the safety of humans, systems, and infrastructures, as well as cybersecurity threats. Safety and security of AI systems are generally defined as Trustworthy AI inherent properties [1]. Numerous software tools exist nowadays to measure quantitatively these properties [2, 3]. Through the Enfield project, the foundation of a non-sectoral methodology is being developed to quantitatively assess safety and security risks of AI systems.

References

- [1] [Ethics guidelines for trustworthy AI](#), High-Level Expert Group on AI, European Commission, 2019.
 [2] Mentzas, Gregoris, et al. "Exploring the landscape of trustworthy artificial intelligence: Status and challenges." *Intelligent Decision Technologies* 18.2: 837-854, 2024.
 [3] Polemi N, Praça I, Kioskli K and Bécue A. "Challenges and efforts in managing AI trustworthiness risks: a state of knowledge.", *Front. Big Data* 7:1381163, 2024.

SCIENTIFIC CHALLENGES

Existing software tools generally quantify and monitor the safety and security of AI systems by computing individual scores related to each property, AI safety and security may encompass, such as reliability and confidentiality. But only a few attempts have been made to adopt a more holistic approach for measuring these properties in a unified and coherent manner [4, 5]. In other words, computing a global AI safety and security score, aggregating different metrics related to AI safety and security properties, remains a challenge.

References

- [4] Huertas Celdran A, Kreischer J, Demirci M, Leupp J, Sánchez Sánchez PM, Figueredo Franco M, et al. A Framework Quantifying Trustworthiness of Supervised Machine and Deep Learning Models. In *SafeAI2023: The AAAI's Workshop on Artificial Intelligence Safety*. 2023.
 [5] Mattioli J, Sohler H, Delaborde A, Amokrane-Ferka K, Awadid A, Chihani Z, et al. An overview of key trustworthiness attributes and KPIs for trusted ML-based systems engineering. *AI and Ethics*. 2024.

RESEARCH ACTIVITIES

To address this challenge, the selected candidate should perform the following research activities:

- **State-of-the-art regarding computable metrics related to safety and security of AI systems**
 The Enfield risk assessment framework is based on existing metrics. The candidate should eventually update it and identify the gaps, where metrics are eventually missing and need to be developed. This possible new development may or may not be part of the work performed through this challenge.
- **State-of-the-art regarding the measurement of AI safety and security properties in an integrated and unified manner**
 The candidate should produce a state-of-the-art regarding the computation of a global AI safety and security score, aggregating different metrics related to AI safety and security properties.
- **Measurement of AI safety and security properties in an integrated and unified manner supporting the Enfield risk assessment framework**
 Based on the aforementioned state-of-the-art and the Enfield risk assessment framework, the candidate should develop a method to compute a global AI safety and security score, aggregating the metrics related to AI safety and security used by the Enfield risk assessment framework. This method will be

tested thanks to the different sectoral scenarios provided by the ENFIELD consortium.

EXPECTED RESULTS

Through the ENFIELD project, the foundation of a non-sectoral methodology is being developed to quantitatively assess safety and security risks of AI systems thanks to existing and new software tools and computable metrics. The work conducted through this challenge should support the development and improvement of this methodology, and its implementation across the different sectoral scenarios provided by the ENFIELD consortium.

This work is expected to result in at least one peer-reviewed scientific publication. The code developed during the project should be publicly made available through a Github repository, or similar platform.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[DTI](#) - Danish Technological Institute / François Picard

[NTNU](#) / Georgios Spathoulas and Georgios Kavallieratos

Research Pillar: Trustworthy AI

T-AI.7 | LLM Safety and Security

Keywords: LLM Safety; LLM Security; Adversarial Robustness; Alignment; Prompt Injection

STATE-OF-THE-ART

LLM safety centres on the responsible development and deployment of models to avoid unintended or malicious harms. Four key areas of concern are **value misalignment** (social bias, privacy leakage, toxicity, ethical violations), **misuse** (e.g., misinformation, deepfakes), **autonomous AI risks**, and robustness to **adversarial attacks** [see arxiv.org/abs/2412.17686]. Regarding the latter, persistent vulnerabilities are frequent in the LLMs available today: jailbreaking attacks can be mounted in both black-box and white-box settings to bypass safeguards, and red-teaming, both manual and automated, has exposed gaps in current filter-and-defend pipelines. Proposed defences include both external safeguards (e.g., gradient-based jailbreak detection and safety reminder prompts) and internal protections (like adversarial tuning). Nevertheless, obfuscated attacks, potentially over a longer multi-turn dialogue, remain hard to catch while preventing over-defensiveness. Integrating emerging interpretability techniques could provide a way towards enhancing LLM safety. Furthermore, as LLMs increasingly drive autonomous agents, with capabilities for planning, tool use, and environment interaction, risks like goal misalignment and unpredictable multi-agent dynamics highlight the need for dedicated safety and security frameworks. Broadly, the domain of **LLM security** is concerned with protecting models and their infrastructure from technical threats that could compromise confidentiality, integrity, or availability. One interesting approach takes a systems perspective [arxiv.org/html/2402.18649] – formulating LLM systems security as constraints on probabilistic information-flow across system components (core models, sandboxes, plugins, and frontends) – finding that context-aware, multi-layer defences are crucial.

SCIENTIFIC CHALLENGES

We invite proposals addressing one or more of the issues introduced above, for example:

- **Adaptive defences:** Building LLM defences that learn and evolve over extended dialogues to catch stealthy, obfuscated attacks without over-censoring
- **Interpretability-guided defences:** Leveraging LLM interpretability techniques (e.g., circuit-level concept analysis, attribution-guided pruning) to pinpoint internal pathways of unsafe behaviors and enable precise interventions
- **Domain-specific safety adaptations:** Tailoring LLM safety methods to industry needs, e.g., for applications in the ENFIELD verticals (energy, healthcare, manufacturing, space) and critical digital infrastructure, where regulatory, ethical, or real-time constraints impose unique safety requirements
- **Lightweight robustness enhancement:** Developing efficient, principled methods to improve LLM behaviour under adversarial or malformed inputs without full adversarial training
- **Resource-constrained alignment:** Exploring alignment and steering techniques (e.g., parameter-efficient fine-tuning, retrieval-based safety layers) that work with limited compute budgets typical of edge or domain-specific deployments
- **Continuous monitoring and self-healing:** Investigating lightweight runtime monitoring agents and self-corrective modules that detect and mitigate emerging safety/security breaches in deployed LLM systems

Systems-level security frameworks: Extending frameworks for information flow in LLM systems (incl. LLM agents) to practical, context-aware enforcement across system components

RESEARCH ACTIVITIES

Relevant research activities could include:

- **Landscape mapping:** Conducting a targeted literature review of existing approaches, potentially highlighting gaps in compositional or domain-specific settings

- Prototype design and implementation: Developing proof-of-concept modules that plug into existing LLM systems or agents
- Method or framework design: Formalising defence approaches – defining core algorithms, theoretical guarantees, or integration strategies – leveraging new techniques or applying and extending existing ideas from the literature in new domains
- Benchmarking and applied case studies: Creating or extending attack scenarios, evaluating proposed methods or prototypes on representative use cases in ENFIELD verticals, using open or synthetic datasets

EXPECTED RESULTS

Valuable contributions from the exchange can for example include:

- Novel defence mechanisms
- Safety/security frameworks and evaluation tools
- Open benchmark suites
- Applied prototypes
- Dissemination materials and activities

At least one peer-reviewed publication at an international academic venue or in a high-impact journal is expected from the exchange

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[TELENOR](#) / Jeriek Van den Abeele

[NTNU](#) / Georgios Spathoulas and Georgios Kavallieratos

Industrial Domain: Energy

IDE.1 | AI for Green Energy curtailment minimization

Keywords: energy; curtailment; prediction; llms

STATE-OF-THE-ART

Energy technical curtailment is the deliberate reduction of energy output from renewable sources to maintain grid stability. It has been exacerbated by increased penetration of wind and solar power, resulting in grid congestion and oversupply issues. Curtailment is triggered by the Grid operator and affects mostly Renewable producers.

Accurate **forecasting** of both energy generation and demand along with Power Plant Control enable solar and wind farms to dynamically respond to market signals and grid conditions, by actuating on the units' output. However, Renewable energy is still curtailed which affects Return-on-Investment (RoI) of solar and wind farms. Forecast AI tools combined with other set of instruments such as Time-of-use tariffs or short size storage (batteries) have been helping workaround the problem. Nevertheless, the specific context of each Renewable promotor may require a customized solution to mitigate and minimize the amount of curtailed Renewable Energy.

SCIENTIFIC CHALLENGES

- Infer from grid and weather features, if conditions are prone to technical curtailment of renewable energy
- The bid (amount of energy sold and price) forecast needs to account the risk of technical curtailment while maximizing RoI
- For each context of a Renewable promotor a set of solutions for the use of potential curtailed energy should be ranked. These solutions may require LLMs application regarding interaction with user for the best tech-economic fit.

RESEARCH ACTIVITIES

The proponent should develop the two AI models above mentioned (prediction and generative) addressing the scientific challenges and determining the features needed to feed the models.

Both models should comply with trustworthy definition and application to the models as stated within ENFIELD project.

EXPECTED RESULTS

The project is expected to produce an application that:

- Will have graphical user interface under a web browser.
- The application will input the generation site features and will automatically acquire data from the grid operator, weather data suppliers and other third-party publicly available data to run the models
- The first model should produce a bid forecast (amount of energy to sell and at what price), considering locally the risk of technical curtailment

The second model should be triggered (appearing as a user interface) if technical curtailment conditions show a high probability (>60%). This should interact with the user through a chat box, suggesting potential uses for the curtailed energy and refining the recommendations on each interaction until alternatives are exhausted or the user is satisfied.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[EDP CNET](#) / Manuel Pio da Silva

Industrial Domain: Energy

IDE.2 | Explainable fault detection and classification in power grids

Keywords: Spatial-temporal data, explainability, graph neural networks, fault detection and classification

STATE-OF-THE-ART

Fast, automated explanation of power system events is increasingly important for grid resilience and operator awareness. Modern techniques rely on analyzing frequency and voltage signals to localize faults, assess their impact, and identify the most affected components. Signal-based methods use multivariate time series and voltage/frequency heatmaps to detect and explain transient events. Graph-based models, such as Gated Graph Neural Networks², leverage the grid topology to improve fault localization and causal inference. To enhance interpretability, domain knowledge should be incorporated into the reports, such as generator inertia, short-circuit ratio, and historical load-shedding patterns, and currently, this is a gap in the state-of-the-art. These contextual features help explain why certain generators or zones were more affected, beyond mere proximity to faults. Recent work³ highlights AI-driven visualization techniques that make system dynamics “visible,” aiding explainability. One possibility for communicating this information is via text generation modules that translate signal data into human-readable summaries, describing what happened, where, and why. A human-in-the-loop process allows iterative refinement, where experts rate explanation quality and guide model improvement.

SCIENTIFIC CHALLENGES

The core scientific challenge is to automatically generate accurate and human-readable reports that explain power system events, such as short circuits or generation losses, based solely on measured frequency and voltage signals. This requires detecting the event location, identifying affected areas and generators, and understanding protection actions like load shedding. The solution must integrate time-series analysis with domain-specific knowledge (e.g., inertia), and translate complex multivariate data into natural language summaries. Ensuring interpretability, generalization across scenarios, and iterative improvement through expert feedback adds further complexity. Balancing technical accuracy, speed, and clarity in a dynamic grid environment is a key research hurdle.

RESEARCH ACTIVITIES

The research activities focus on developing a pipeline that transforms power system measurements into clear, explanatory reports.

This includes (1) designing algorithms to detect and localize events using multivariate time series and heatmaps; (2) integrating domain knowledge, such as generator inertia, short-circuit ratio, and protection locations, into the interpretation process; and (3) creating natural language generation models to produce textual explanations. A human-in-the-loop approach is used to refine outputs based on expert feedback. Simulations on a modified IEEE 39-bus system provide realistic scenarios, including short circuits and generation losses, for testing. Validation includes both expert review and general user evaluation to assess the clarity and usefulness of the generated reports.

EXPECTED RESULTS

The expected results include an automated system capable of generating fast, accurate, and understandable reports of power system events based on measured signals. It will detect event types, locations, affected zones, and key generators or protection actions, while incorporating relevant domain knowledge. A publication and software code is expected as final outcome from this work.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[INESC TEC](#) / Dr. Francisco Fernandes

² de Freitas, J. T., Coelho, F. G. F. (2021). Fault localization method for power distribution systems based on gated graph neural networks. *Electrical Engineering*, 103(5), 2259-2266.

³ Miranda, V., Cardoso, P. A., Bessa, R. J., Decker, I. (2019). Through the looking glass: Seeing events in power systems dynamics. *International Journal of Electrical Power & Energy Systems*, 106, 411-419.

Industrial Domain: Energy

IDE.3 | Combine AI with LLM for clear human interaction with complex data

Keywords: Human-AI interaction, Large Language Models, Green AI, Energy Industry

STATE-OF-THE-ART

Versatile AI-based algorithms for data interpretation are emerging but remain narrowly oriented, tackling specific, well-defined problems. Tools such as the SMASH platform (Tom Wilcox, 2019) enable data storage, querying, and visualization but still rely heavily on human domain experts for interpreting and presenting results. Large Language Models (LLMs) have been applied to various domains, demonstrating potential in reducing this reliance. For example, LLMs have improved the annotation of time-series data (Mi Zhou, 2023), automated domain-specific data curation through SEED (Zui Chen, 2023), and enhanced interpretability with rule-based guidance for forecasting (Dilini Rajapaksha, 2022). LLM-based Multi-Agent Systems (MAS) extend these capabilities further by orchestrating specialized agents to collaboratively solve complex tasks. Notable examples include MARCO (Shrimal et al., 2024), which manages user-agent interactions through planning and reasoning agents, and ControlAgent (Guo et al., 2024), which integrates LLMs for automating control system design. The L2MAC framework (Jin et al., 2024) demonstrates MAS's ability to dynamically manage memory and execution context for large-scale code generation. These systems showcase how LLMs and MAS frameworks can deliver state-of-the-art performance in tasks like code generation, question answering, and data visualization.

By building on these advances, the proposed project aims to create a MAS with orchestrator, coding, and interpretation agents. This system will simplify user interactions with complex data, leveraging the strengths of LLMs and MAS to achieve adaptive and intuitive performance.

References

- Haolin Jin, Linghan Huang, Haipeng Cai, Jun Yan, Bo Li, Huaming Chen. From LLMs to LLM-based Agents for Software Engineering: A Survey of Current, Challenges and Future. 2024. arXiv.org.
- Junwei Liu, Kaixin Wang, Yixuan Chen, Xin Peng, Zhenpeng Chen, Lingming Zhang et al. Large Language Model-Based Agents for Software Engineering: A Survey. 2024. arXiv.org.
- Xing-ming Guo, Darioush Keivan, U. Syed, Lianhui Qin, Huan Zhang, G. Dullerud et al. ControlAgent: Automating Control System Design via Novel Integration of LLM Agents and Domain Expertise. 2024. arXiv.org.
- Anubhav Shrimal, Stanley Kanagaraj, Kriti Biswas, Swarnalatha Raghuraman, Anish Nediyanath, Yi Zhang et al. MARCO: Multi-Agent Real-time Chat Orchestration. 2024. Conference on Empirical Methods in Natural Language Processing.
- Dilini Rajapaksha, C. B. (2022). LIMREF: Local Interpretable Model Agnostic Rule-based Explanations for Forecasting, with an Application to Electricity Smart Meter Data. The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22), (str. 12098-12107).
- Jiaqi Ruan, G. L. (2023). Applying Large Language Models to Power Systems: Potential Security Threats. arXiv, 2311.13361.
- Mi Zhou, F. L. (2023). Meta In-Context Learning: Harnessing Large Language Models for Electrical Data Classification. Energies, MDPI, Volume 16(18), 1-18.
- Muhammad Usman Hadi, Q. A.-T. (2023). Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. TechRxiv.
- Qing Lyu, J. T. (2023). Translating Radiology Reports into Plain Language using ChatGPT and GPT-4 with Prompt Learning: Promising Results, Limitations, and Potential. Visual computing for industry, biomedicine, and art, 6(1), 2303.09038.
- Tom Wilcox, N. J. (2019). A Big Data platform for smart meter data analytics. Computers in Industry, Volume 105, 250-259.
- Zui Chen, L. C. (2023). SEED: Domain-Specific Data Curation With Large Language Models. arXiv e-prints, arXiv:2310.00749.

SCIENTIFIC CHALLENGES

Modern devices can be considered as data generators, either utilizing the generated data directly for their current tasks or storing the data for potential future use. Smart energy meters and other smart devices, e.g. heat pumps, EV chargers, PVs and inverters all fall into this category. Smart energy meters can generate a snapshot of most main electrical parameters periodically, down to a one second interval, and standard AMI energy meters usually generate data on 15-minute intervals. This enormous amount of complex and, at first sight, uncorrelated smart meter and other smart device data is generated at multiple collection points. The data holds various information about the observed energy system but is usually hidden from direct interpretation. Finding patterns in this data pool and interpreting it therefore requires expert knowledge of energy systems and their real-world implementations. It is nearly impossible to query the data without deep domain knowledge and specifically developed statistics and AI SW tools. Extracting derivatives in form of co-dependencies, correlation and other from the source data requires additional knowledge of the data and the tools being used. The topic falls under the Green AI (some preprocessing could be conducted on smart meter Edge devices) and Human-centric AI (human prompt and understanding friendly interface) Pillars.

RESEARCH ACTIVITIES

The use-case (UC) focuses on simplifying and clarifying the interpretability of complex, energy sector related data. The Green transition brings many new concepts to residential as well as industrial users, who will have to start working with, till now, unimportant data, connected to energy generation and consumption. Understanding the values, their interconnection and context will present an important factor in navigating the energy ecosystem. This task currently requires a domain expert,

who has vast experience with handling smart meter data. This UC aims to develop a solution which will i) accept human prompts for interacting with data, ii) prepare necessary data processing algorithms to extract the required data from complex data pools, iii) analyze the outcome and formulate the result in form understandable for the user. The main goal is to research into LLM multi-agent system (MAS) and develop a framework to facilitate intuitive interaction with data using natural language prompts consisting in:

- An orchestrator agent to decompose tasks, assign roles, and manage workflows.
- A coding agent to generate database queries and visualization scripts.
- An interpretation agent to analyze results and present them in a clear and actionable format.

The resulting system will empower users to interact with and extract insights from complex datasets, meeting the increasing demand for user-friendly AI solutions across industries.

EXPECTED RESULTS

This project aims to develop a multi-agent system (MAS) that integrates large language models (LLMs), task-specific AI agents and tools to handle complex user queries and interactions with large datasets. It will include: 1) an orchestrator agent to dynamically allocate tasks, 2) coding agent to generate database queries and visualization scripts, 3) interpretation agent to analyze and summarize data results. The expected outcome is a fully functional MAS capable of simplifying interactions with complex data. Key Performance Indicators (KPIs): **Response Time, Clarity of Results, Task Accuracy, Context Awareness, Scalability.**

AI requirements: The proposed system will employ a multi-tiered approach to AI integration. Use of 1) high-level LLM which serves as the orchestrator agent to translate human-friendly prompts 2) low-level LLM that work on smart meter-generated data to produce the queried data 3) high-level LLM as interpretation role, which presents the data in a clear and easily understandable form to the user. The presented AI requirements are aligned with Human-centric AI.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[Know Center Research GmbH](#) / Priv.-Doz. Dr.techn. Eduardo Veas, MSc

Industrial Domain: Energy

IDE.4 | LLM based data queries and visualization

Keywords: Large Language Model, Human-centric AI, Green AI, Energy Industry

STATE-OF-THE-ART

Language models (LLMs) have been increasingly used in various applications, including coding, and visual reasoning. Early systems like SMASH (Tom Wilcox, 2019) focused on data storage, querying, and visualization but relied heavily on human domain experts for interpretation. LLMs have since advanced these capabilities by enabling automated data classification (Mi Zhou, 2023), reducing the need for annotated datasets, and enhancing decision-making with interpretable guidance frameworks (Dilini Rajapaksha, 2022). Recent advancements in LLM-based multi-agent systems have brought significant improvements in generating queries to databases from free text and generating code for visualizing the outcome. HYDRA (Ke et al., 2024) dynamically coordinates agents for compositional reasoning and tailored visualization code generation, while MAC-SQL (Wang et al., 2023) uses collaborative agents for Text-to-SQL tasks by decomposing queries and refining SQL commands. Self-Organized Agents (Ishibashi et al., 2024) focus on ultra-large-scale code generation, achieving higher accuracy through agent collaboration. CodeRefine (Trofimova et al., 2024) enhances LLM-generated implementations with task-aware vector embeddings and refined code synthesis.

By leveraging LLMs' ability to parse natural language and integrate domain-specific tools it is possible to tackle challenges like dynamic query formulation, data visualization, and collaborative code refinement. MAC-SQL and SoA highlight advances in query decomposition and scalable code generation. However, further research is needed to fully exploit MAS in diverse applications.

References

- Bing Wang, Changyu Ren, Jian Yang, Xinnian Liang, Jiaqi Bai, Qian-Wen Zhang et al. MAC-SQL: A Multi-Agent Collaborative Framework for Text-to-SQL. 2023. arXiv.org.
- Fucai Ke, Zhixi Cai, Simindokht Jahangard, Weiqing Wang, P. D. Haghighi, Hamid Reza Tofighi. HYDRA: A Hyper Agent for Dynamic Compositional Visual Reasoning. 2024. arXiv.org.
- Yoichi Ishibashi, Yoshimasa Nishimura. Self-Organized Agents: A LLM Multi-Agent Framework toward Ultra Large-Scale Code Generation and Optimization. 2024. arXiv.org.
- Ekaterina Trofimova, Emil Sataev, Abhijit Singh Jowhari. CodeRefine: A Pipeline for Enhancing LLM-Generated Code Implementations of Research Papers. 2024. arXiv.org.
- Dilini Rajapaksha, C. B. (2022). LIMREF: Local Interpretable Model Agnostic Rule-based Explanations for Forecasting, with an Application to Electricity Smart Meter Data. The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22), (str: 12098-12107).
- Jiaqi Ruan, G. L. (2023). Applying Large Language Models to Power Systems: Potential Security Threats. arXiv, 2311.13361.
- Mi Zhou, F. L. (2023). Meta In-Context Learning: Harnessing Large Language Models for Electrical Data Classification. Energies, MDPI, Volume 16(18), 1-18.
- Muhammad Usman Hadi, Q. A.-T. (2023). Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. TechRxiv.
- Qing Lyu, J. T. (2023). Translating Radiology Reports into Plain Language using ChatGPT and GPT-4 with Prompt Learning: Promising Results, Limitations, and Potential. Visual computing for industry, biomedicine, and art, 6(1), 2303.09038.
- Tom Wilcox, N. J. (2019). A Big Data platform for smart meter data analytics. Computers in Industry, Volume 105, 250-259.
- Zui Chen, L. C. (2023). SEED: Domain-Specific Data Curation With Large Language Models. arXiv e-prints, arXiv:2310.00749.

SCIENTIFIC CHALLENGES

Modern smart devices, including energy meters, heat pumps, EV chargers, photovoltaic (PV) systems, and inverters, are prolific generators of data. This data is either utilized in real time to support ongoing operations or stored for subsequent analysis and decision-making. Smart energy meters, for instance, are capable of capturing detailed snapshots of electrical parameters at intervals as short as one second, while Advanced Metering Infrastructure (AMI) meters typically operate at 15-minute intervals. This results in the accumulation of vast, complex datasets, often generated across multiple collection points. These datasets, while rich in information, present significant interpretive challenges. Raw data often lacks obvious correlations or patterns, obscuring its potential to provide actionable insights. Extracting meaningful derivatives, such as dependencies, correlations, and system-wide trends, necessitates both domain-specific knowledge of energy systems and proficiency with analytical tools. This challenge underscores the importance of aligning the data interpretation process with the principles of Green AI, emphasizing efficiency through preprocessing at the edge device level, and Human-centric AI, which prioritizes the development of intuitive, user-friendly interfaces. Such an approach aims to democratize access to energy data insights, enabling users to interact effectively with these complex datasets without requiring specialized expertise.

RESEARCH ACTIVITIES

The use-case (UC) aims to simplify and enhance the interpretability of complex energy-sector data. The green transition introduces numerous new concepts to both residential as well as industrial users, requiring them to engage with previously overlooked data related to energy generation and consumption. Understanding the interconnections and contextual relevance of this data is crucial for navigating the evolving energy ecosystem. Currently, this process relies heavily on domain experts with significant experience in handling smart meter data. The objective is to research and develop a solution that can:

i) generate code for querying relevant data from complex datasets, ii) create code to produce visualizations from the query results. The proposed system will empower users to interact with and derive insights from intricate datasets, addressing the growing need for user-friendly AI solutions across various industries.

EXPECTED RESULTS

This project aims to develop LLM based solution, to generate queries to databases from free text and to create visualizations with the outcomes. The expected outcome is a set of LLM to successfully handle complex queries as well as for the generation of the code needs to visualize the resulting data. Key Performance Indicators (KPIs): **Task Accuracy, Context Awareness, Scalability.**

AI requirements: Use of 1) low-level LLM that work on smart meter-generated data to produce the queried data. The results are delivered to the high-level LLM, which illustrates the data in visualizations or a dashboard for the user. The presented AI requirements are aligned with Human-centric AI.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[Know Center Research GmbH](#) / Priv.-Doz. Dr.techn. Eduardo Veas, MSc

Industrial Domain: Manufacturing

IDM.1 | Adversarial Robustness Metrics and Evaluation for AI-Driven Safety Systems

Keywords: Adversarial robustness, risk assessment, safety-critical AI, threat detection, performance metrics

STATE-OF-THE-ART

The recent EU Machinery Regulation [1] opens for the development and the use of safety equipment that employs self-evolving behaviour, including AI-driven processes. In a manufacturing context, AI-driven safety sensors may be used to ensure the safety of operators evolving near robotics systems, or any other machinery. According to the EU Machinery Regulation, such safety sensors should be robust against hazardous situations, such as adversarial attacks. Several conceptual frameworks have been developed to assess this kind of cybersecurity risks related to AI systems [2, 3, 4].

References

- [1] [EU Machinery Regulation](#), European Commission, 2023
- [2] [Framework for Artificial Intelligence Cybersecurity Practices from ENISA](#)
- [3] [NIST's taxonomy & terminology related to adversarial machine learning](#)
- [4] [MITRE's Adversarial Threat Landscape for Artificial-Intelligence Systems](#)

SCIENTIFIC CHALLENGES

- **Definition of adversarial robustness metrics:** New quantitative metrics must be developed, measuring accurately the robustness of AI models against adversarial attacks.
- **Robustness score thresholds considering real-world manufacturing safety requirements and plausible threat scenarios:** The adversarial robustness scores computed through dedicated metrics must be evaluated considering real-world manufacturing safety requirements and plausible threat scenarios.
- **Monitoring and evaluation of adversarial attack detection:** There is a need for implementing testing protocols for evaluating the effectiveness and reliability of adversarial attack detection integrated in AI-driven safety equipment.

RESEARCH ACTIVITIES

To address this challenge, the selected candidate should perform the following research activities:

- **Development of adversarial robustness metrics for AI-driven safety systems**
The candidate should investigate and develop new quantitative metrics to measure the robustness of AI-driven safety equipment against adversarial attacks.
- **Contextual adversarial robustness threshold**
The candidate should develop a methodology to evaluate proper thresholds related to the aforementioned adversarial robustness metrics considering real-world manufacturing safety requirements and plausible threat scenarios.
- **Adversarial attack detection and testing protocols**
The candidate should design attack detection methods and devise comprehensive testing protocols related to them. These methods and protocols should demonstrate the performance of the AI-driven safety systems against adversarial impacts, as well as the impact of attack detections on the overall system effectiveness and reliability.

EXPECTED RESULTS

Practically speaking, the candidate will work with AI-driven safety cameras and on real-world manufacturing scenarios provided by the host institution and more generally the Enfield consortium. The candidate will work with public datasets or self-provided ones.

The expected results are the following:

- Development of a well-defined metric, or set of metrics, for quantifying adversarial robustness, suitable for regulatory and industrial use in AI-driven safety systems.

- Development of a methodology and practical recommendations for setting adversarial robustness thresholds considering real-world manufacturing safety requirements and plausible threat scenarios.
- Design and development of attack detection methods and testing protocols for demonstrating the performance of the AI-driven safety systems against adversarial impacts and attack detection impact on the overall system effectiveness and reliability.

This work is expected to result in at least one peer-reviewed scientific publication. The methodologies and code developed during the project should be publicly made available.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[DTI](#) - Danish Technological Institute / François Picard

[POLIMI](#) / Walter Quadri

[PREDICT](#) / Leïla Belmerhnia

Industrial Domain: Manufacturing

IDM.2 | Bias-aware People Detection: Quantifying and Mitigating Dataset and Model Biases Towards Safety Certification

Keywords: People detection, bias, Mean Time to Dangerous Failure (MTTFd), safety

STATE-OF-THE-ART

The recent EU Machinery Regulation [1] opens for the development and the use of safety equipment that employs self-evolving behaviour, including AI-driven processes. This regulation goes as well along with specific safety standards, such as ISO 13849-1:2023 [2] related to Mean Time to Dangerous Failure (MTTFd) a crucial reliability metric used to assess the safety performance of a system or component, particularly in safety-critical applications.

In a manufacturing context, safety cameras with AI functionalities to detect people are more and more considered to ensure the reliable monitoring and the safety of operators evolving near robotics systems, or any other machinery. Safety cameras with AI functionalities must be certified according to the EU Machinery Regulation and thanks to safety standards, such as ISO 13849-1:2023.

References

- [1] [EU Machinery Regulation](#), European Commission, 2023
 [2] [ISO 13849-1:2023](#), Safety of machinery — Safety-related parts of control systems, 2023.

SCIENTIFIC CHALLENGES

- Quantification of bias related to AI model and training dataset**
 Detecting people with AI for safety concern has proven to be highly biased, not only relatively to gender and race [3] but also posture [4]. Quantifying these biases and their impact on the AI system performance remains a challenge.
- Minimum acceptable variance in class representation**
 It is still today difficult to measure the minimum acceptable variance in class representation that allows near-equal detection performance and reliability for all represented classes in the training dataset. This is particularly the case when considering safety equipment such as AI-driven safety cameras.
- Towards safety certification**
 Relating bias quantification measures to safety-critical performance indicators, such as MTTFd, still poses a significant challenge.

References

- [3] D. Zhao, A. Wang and O. Russakovsky, "Understanding and Evaluating Racial Biases in Image Captioning," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 14810-14820
 [4] Huang, Y., Sun, B., Kan, H., Zhuang, J., Qin, Z. (2019). FollowMeUp Sports: New Benchmark for 2D Human Keypoint Recognition. In: Lin, Z., et al. Pattern Recognition and Computer Vision. PRCV 2019. Lecture Notes in Computer Science, vol 11859. Springer, Cham.

RESEARCH ACTIVITIES

To address this challenge, the selected candidate should perform the following research activities:

- Development of bias quantification metrics for AI-driven safety cameras**
 The candidate should investigate and develop new bias quantitative metrics when considering AI models and datasets used for people detection in a manufacturing and safety context.
- Analysis of AI-driven safety camera performance considering bias**
 Based on the aforementioned bias quantification metrics, the candidate should analyse the AI-driven safety camera performance and reliability. The candidate should especially focus on the difference in class representation (considering gender, race and posture) and its impact on the overall system performance and reliability. The goal is to link this difference with the system's performance and reliability for all represented classes in the training dataset.
- Mapping between bias measurements and MTTFd**

The candidate should study the relation between bias measurements with MTTFd, considering the overall safety system performance and reliability.

EXPECTED RESULTS

Practically speaking, the candidate will work with AI-driven safety cameras and on real-world manufacturing scenarios provided by the host institution and more generally the Enfield consortium. The candidate will work with public datasets or self-provided ones.

The expected results are the following:

- Development of bias quantification metrics, related to AI models and datasets used for people detection in a manufacturing and safety context.
- Benchmark demonstrating the relationship between class representation, bias, and safety performance and reliability. This benchmark should allow the quantification of the minimum acceptable variance in class representation that allows near-equal detection performance and reliability for all represented classes in the training dataset.
- Mapping of bias quantification to MTTFd.

This work is expected to result in at least one peer-reviewed scientific publication. The methodologies, datasets, AI models, and code developed during the project should be publicly made available.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[DTI](#) - Danish Technological Institute / François Picard
[PREDICT](#) / Leïla Belmerhnia

Industrial Domain: Manufacturing

IDM.3 | Explainable Uncertainty-Aware Decision-Making for AI Systems in Manufacturing

Keywords: Uncertainty quantification, explainable AI, safety-critical systems, Bayesian machine learning

STATE-OF-THE-ART

It is well-known that deep neural networks are poorly calibrated and produce poor estimates with very high confidence when operating outside of the training distribution [1,4]. This remains a challenge when it comes to safety-critical systems, such as robotics systems.

Various methods for uncertainty quantification in deep neural networks have been proposed, e.g., Bayesian Neural Networks, ensemble methods, sampling-based (MCMC) methods, etc. [2]. Empirically, ensemble methods are found to be some of the best-calibrated uncertainty quantification methods but come at higher computational cost [3]. Uncertainty quantification (UQ) is important not only for safety, but also for more data-efficient training, as knowing the regions of highest uncertainty in the model state-space makes it possible to use active learning [6,7].

References

- [1] Wang, Cheng. "Calibration in deep learning: A survey of the state-of-the-art." arXiv preprint arXiv:2308.01222 (2023).
- [2] Gawlikowski, Jakob, et al. "A survey of uncertainty in deep neural networks." Artificial Intelligence Review 56.Suppl 1 (2023): 1513-1589.
- [3] Ovadia, Yaniv, et al. "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift." Advances in neural information processing systems 32 (2019).
- [4] Mena, José, Oriol Pujol, and Jordi Vitrià. "A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective." ACM Computing Surveys (CSUR) 54.9 (2021): 1-35.
- [5] Billard, Aude, et al. "A roadmap for AI in robotics." Nature Machine Intelligence (2025): 1-7.
- [6] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, "A survey of deep active learning," ACM computing surveys (CSUR), vol. 54, no. 9, pp. 1–40, 2021.
- [7] D. Li, Z. Wang, Y. Chen, R. Jiang, W. Ding, and M. Okumura, "A survey on deep active learning: Recent advances and new frontiers," IEEE Transactions on Neural Networks and Learning Systems, 2024.

SCIENTIFIC CHALLENGES

The scientific challenges related to UQ in Manufacturing are the following:

- Defining real-world benchmarks for uncertainty quantification frameworks
- Defining ground truth uncertainty measures for a given problem
- Developing methods that can explain the reason for a given uncertainty measure
- Developing user-friendly human-machine interaction frameworks that can communicate the meaning of uncertainty in a task-specific way to non-experts

RESEARCH ACTIVITIES

To address this challenge, the selected candidate should perform the following research activities:

- Collecting dataset for uncertainty quantification
- Creating benchmark for uncertainty quantification based on real-world use case
- Investigating methods for UQ
- Investigating methods for adding explainability to UQ
- Developing a framework for training explainable uncertainty-aware AI models
- Preparing of scientific manuscripts for conference and journal publications

EXPECTED RESULTS

The expected results are the following:

- A framework for uncertainty quantification in AI models
- A demonstrator
- At least 1 conference and 1 journal publication

The candidate will work with public datasets or self-provided ones. The methodologies, datasets, AI models, and code developed during the project should be publicly made available.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[DTI](#) - Danish Technological Institute / François Picard, [POLIMI](#) / Walter Quadrini, [PREDICT](#) / Leïla Belmerhnia

Industrial Domain: Manufacturing

IDM-A.AI.4 | Multimodal AI for Human-State Monitoring

Keywords: multimodal learning; wearable sensors; privacy; human-centric manufacturing; Adaptive-AI

STATE-OF-THE-ART

The ongoing ageing of the European population is expected to significantly impact the manufacturing sector, where operators will, on average, no longer be suited for repetitive or physically demanding tasks. Current mitigation strategies include the use of empowering technologies, such as exoskeletons and collaborative robots, which, however, tend to increase the lead time of operations. There is therefore a growing need for systems that are activated only under conditions of physical, mental, or emotional stress, striking a fair balance between operator well-being and overall production performance. Several technologies originating from the healthcare domain have been applied in various sectors to evaluate the stress levels of individuals performing tasks. However, the manufacturing environment presents unique challenges, such as manual labor, long shifts, and high operator mobility, that render many existing experimental setups and stress estimation techniques, originally designed for medical patients or office workers, unsuitable for this context.

To address this, Adaptive AI approaches can be explored to enable personalized, context-aware assistance that continuously adjusts to individual operators and dynamic shop floor conditions. Such models could support selective activation of support systems based on evolving patterns of stress, fatigue, or task difficulty, ensuring both responsiveness and robustness without manual recalibration.

SCIENTIFIC CHALLENGES

The hosting institution will provide access to a wide range of wearable sensing equipment, including EEG, ECG, EMG, eye-trackers, electrodermal activity (EDA), skin temperature (ST) sensors, and cameras. These sources may be selectively integrated into the data collection pipeline, depending on the research focus. The selected project is expected to develop and validate a multimodal model capable of detecting operators' stress (physical, mental, or emotional) using the available sensor data. Support for sensor integration and system setup will be provided by the host institution.

The researcher will be responsible for leading the scientific and implementation activities, which are expected to include:

- Knowledge extraction about the physical, mental, or emotional stress from the data available
- Addressing privacy-related issues in framework of the EU regulations.
- Demonstration in a laboratory environment with real-like manufacturing tasks (e.g., assembly, kitting, sorting...)

Another key challenge is to develop Adaptive AI techniques that enable models to personalize to different operators and adapt to changing tasks or conditions without full retraining. Approaches may include continual learning, online adaptation, or meta-learning, aiming to ensure long-term robustness and flexibility.

RESEARCH ACTIVITIES

The candidate is expected to have a good knowledge of Multimodal Analysis and Data Processing Methods. Moreover, the following steps should be taken:

- Investigating existing approaches and proposing a novel framework for physical, mental, or emotional stress detection and classification.
- Designing and implementing models capable of estimating physical, mental, or emotional stress in a person involved in manufacturing operations.
- Evaluating the robustness of the proposed framework across diverse operator profiles, (e.g., sex and age).

Where possible, the candidate is encouraged to explore adaptive AI strategies that allow the model to personalize to individual operators and adjust to dynamic task conditions over time.

EXPECTED RESULTS

The outcomes of this work are expected to contribute to improving the well-being and working conditions of operators in manufacturing environments, while also supporting applied research efforts aimed at extending the employability of the European workforce.

The developed model should be released under Open Access to promote transparency and reusability within the research and industrial communities.

The beneficiary is expected to deliver a functional demonstration of the proposed system, as well as produce at least one peer-reviewed journal publication. A detailed publication plan should be included in the proposal and will be evaluated in terms of ambition and feasibility.

An experimental protocol, including a clear data management plan, is strongly encouraged. Alternatively, the candidate may choose to use public datasets, which must be clearly stated and justified.

Where applicable, the candidate is encouraged to incorporate Adaptive AI techniques that enable the system to personalize stress detection models to individual operators and adjust to changes in task context or physiological response over time. This may include, but is not limited to, methods such as online learning, domain adaptation, continual learning, or meta-learning. The goal is to enhance the model's robustness, reduce the need for frequent recalibration, and support long-term deployment in dynamic manufacturing environments.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[POLIMI](#) / Walter Quadrini

Note: This challenge is co-hosted by [IMT Télécom Paris](#)

Industrial Domain: Manufacturing

IDM-A.AI.5 | Egocentric perception for shopfloor operators

Keywords: Egocentric Vision; Multimodal Learning; Manufacturing; Data Processing

STATE-OF-THE-ART

Egocentric vision provides a unique first-person perspective for understanding human activities, serving as a fundamental tool in various domains, including assistive technology [1] and robotics applications [2]. As a result, working with and providing access to such data has become a major trajectory within the research community. Multiple open-source datasets have emerged, such as Ego4D [3] and EPIC-KITCHENS-100 [4], which aim to provide a general understanding of egocentric data and address various theoretical and scientific challenges. However, since most of these datasets focus on general daily life tasks, a major challenge remains: the lack of diversity across different domains—particularly industrial settings.

Furthermore, the emphasis on flexibility in 'Industry 4.0' has resulted in a complex and diverse range of products principles that assembly workers are expected to know. Since this is unrealistic, computer vision-assisted workstations can help by identifying assembly mistakes, reducing the need for rework, and improving operator efficiency. Therefore, it is crucial to develop an industrial-focused egocentric dataset that captures the unique characteristics of manufacturing environments, along with a framework capable of leveraging such data to enhance workplace efficiency.

Currently, in the state of the art, multiple modalities are often used to build a robust understanding and to design assistive frameworks that help operators work efficiently in industrial settings. In this context, as part of a collaboration between POLIMI and IMT, we are planning to create a comprehensive multimodal egocentric dataset that captures complex industrial scenarios. Our goal is to explore the potential of using this dataset to develop an assistive egocentric vision model that integrates multiple modalities – potentially through transformer-based architectures – to build an AI agent capable of understanding and predicting operators' behaviour and supporting them in performing their tasks more effectively.

References

- [1] E. OhnBar, K. Kitani, C. Asakawa, Personalized dynamics models for adaptive assistive navigation systems, in: CORL, 2018.
 [2] H.S. Park, J.-J. Hwang, Y. Niu, J. Shi, Egocentric future localization, in: CVPR, 2016.
 [3] Grauman, Kristen, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger et al. "Ego4d: Around the world in 3,000 hours of egocentric video." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 18995-19012. 2022.
 [4] Damen, Dima, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti et al. "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100." International Journal of Computer Vision (2022): 1-23.

SCIENTIFIC CHALLENGES

The hosting institution will provide access to a data generation environment for egocentric data in the industrial domain.

The researcher will be expected to work effectively with the data collection pipeline established by the hosting and mentoring institution and, depending on the research focus, integrate it into the selected project. The project aims to develop and validate an egocentric model capable of classifying and predicting operators' activities to support the working scenario, using egocentric vision along with other modalities.

The researcher will be responsible for leading the scientific and implementation activities, which are expected to include:

- Fusion and knowledge extraction from multimodal egocentric understanding.
- Addressing privacy-related issues within the framework of EU regulations.
- Demonstrating model performance in a laboratory environment simulating real-world manufacturing tasks (e.g., assembly, kitting, sorting...).

RESEARCH ACTIVITIES

The candidate is expected to have a solid knowledge of Multimodal Data Processing and Computer Vision. Moreover, the following steps should be undertaken:

- Participation in data generation, annotation and cleaning.
- Investigating existing frameworks for using multimodal egocentric data in industrial settings and proposing a novel framework.

- Designing and implementing a new model capable of understanding, classifying, and predicting the activities of operators involved in manufacturing operations.

Evaluating the robustness of the proposed framework using the data generated by POLIMI and IMT.

EXPECTED RESULTS

The following outcomes are expected:

- A new multimodal egocentric dataset for monitoring and understanding operator activity in industrial scenarios. Please note that data generation will be a collaborative effort involving multiple institutions, and active contribution to dataset creation is essential.
- A novel and robust framework that leverages multiple modalities to understand operator activities, ultimately supporting and enhancing efficiency in manufacturing environments.
- At least one peer-reviewed publication for journal publication.
- Cooperation in the paper accompanying the dataset is also expected.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[POLIMI](#) / Walter Quadrini

Note: This challenge is co-hosted by [IMT Télécom Paris](#)

Industrial Domain: Space

IDS.1 | Detection of potential water illegal abstractions using Artificial Intelligence and Earth Observation

Keywords: agriculture, water resources, Artificial Intelligence, Earth Observation, climate crisis

STATE-OF-THE-ART

Remote Sensing and Earth Observation are widely used in irrigation management in order to retrieve useful information about the cultivated crop types and their current crop growth stage. By combining this with evapotranspiration rates, the amount of irrigation needs can be calculated. Furthermore, by monitoring the spatiotemporal patterns of soil moisture using spaceborne sensors the detection of irrigation events can be determined. The employment of Remote Sensing for the detection and monitoring of potential illegal abstractions is still unexplored. Still, a recent research work showed that by monitoring the vegetation health using satellite vegetation indices and by observing unusual patterns of healthy vegetation can be beneficial for the detection of potential illegal water abstractions or illegal irrigation activities. .

SCIENTIFIC CHALLENGES

Agriculture industry around the world suffers from Climate Change. Water scarce areas are suffering from irrigation water shortage due to drought events. Moreover, in areas like Cyprus and the rest of Mediterranean basin is noticed that farmers who are cultivating rainfed crops and not only are illegally using more water for irrigation purposes. The investigation of this research challenge lacks solutions.

RESEARCH ACTIVITIES

- Data collection.
- Literature review
- Development of AI model for detection of illegal water abstractions.
- Preparation of scientific manuscript for journal publication.

EXPECTED RESULTS

- 1 journal publication submitted.
- 1 model/algorithm/software/framework

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[ERATOSTHENES Centre of Excellence](#) / Dr. Michalis Mavrovouniotis and Stelios Neophytides

Industrial Domain: Space

IDS.2 | Causal Machine Learning model to identify agricultural practices aiding in yield productivity improvement using Earth Observation (EO) data

Keywords: machine learning, Earth observation, agriculture, food security, big data, artificial intelligence

STATE-OF-THE-ART

Industrial applications utilizing EO data for yield prediction and estimation in order to support farmers are arising. Causal Machine Learning (CML) is still undiscovered within the discipline of Earth Sciences. CML can help through its capabilities to explore a problem further than correlations of data to a problem by detecting the cause and estimating the cause's effect. Causal Inference and Causal Analysis by employing Double Machine Learning (DML) and other more traditional methodologies are widely used in the sectors of economics and business intelligence.

SCIENTIFIC CHALLENGES

Personalized applications dedicated to farmer's practices in combination with EO data are still an undiscovered path. Farmers must be in the centre to help them improve their yield productivity by identifying the cause effect of the different agricultural practices (e.g., irrigation management, fertilizations) to yield productivity. The combination of causal machine learning on EO is still a not quite explored path.

RESEARCH ACTIVITIES

- Data collection.
- Literature review
- Development of AI model for yield prediction.
- Application of Causal Machine Learning to identify the effect of agricultural practices or environmental conditions to yield.
- Preparation of scientific manuscript for journal publication.

EXPECTED RESULTS

- 1 journal publication submitted.
- 1 model/algorithm/software/framework

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[ERATOSTHENES Centre of Excellence](#) / Dr. Michalis Mavrovouniotis and Stelios Neophytides

Industrial Domain: Space

IDS.3 | Advancing Eelgrass Monitoring Using Stationary High-Altitude Balloons and Hyperspectral AI-based Earth Observations

Keywords: High Altitude Balloons (HABs), hyperspectral imaging, artificial intelligence, eelgrass, Earth observations

STATE-OF-THE-ART

Monitoring coastal vegetation such as eelgrass is critical for assessing ecosystem health, carbon sequestration, and biodiversity in marine environments. Traditional observation techniques rely on in-situ surveys and satellite imagery, but both face significant limitations in spatial, spectral, and temporal resolution, especially in Danish, Baltic and Arctic coastal zones. Stationary High Altitude Balloons (HABs) are a cutting-edge platform, providing sustainable, reusable, high-resolution, and persistent data collection capabilities. Coupling HAB-based hyperspectral imaging with modern AI approaches offers transformative potential for detecting and mapping eelgrass with unprecedented accuracy, surpassing current satellite-based Earth Observation systems like Copernicus.

SCIENTIFIC CHALLENGES

- **Detecting Eelgrass Under Water:** Optical detection of submerged eelgrass is challenged by water column variability, turbidity, and mixed spectral signals.
- **Data Integration:** Combining multi-modal data streams (e.g., RGB, hyperspectral, and AI-generated outputs) for robust monitoring requires novel analysis pipelines.
- **Platform and Payload Integration:** Ensuring the reliable operation of sophisticated imaging instruments on HAB platforms, along with safe and regulatory-compliant launches.
- **Algorithm Development:** AI models capable of robustly classifying eelgrass from hyperspectral data in variable real-world conditions have yet to be validated at scale.

RESEARCH ACTIVITIES

- Design and fly stationary HAB missions equipped with hyperspectral and high-resolution cameras over selected sites.
- Collect and pre-process imaging datasets of eelgrass meadows.
- Develop and train AI algorithms for automated eelgrass segmentation and mapping from balloon-acquired hyperspectral data.
- Compare balloon-based results with satellite and field-based reference data.
- Disseminate findings through conference presentations and peer-reviewed journal submissions.

EXPECTED RESULTS

- High-resolution, georeferenced hyperspectral datasets of Danish, Arctic or Baltic eelgrass meadows from HAB flights.
- Validated AI models or processing pipelines for eelgrass detection and quantification from airborne data.
- 1–2 journal or conference publications.
- Demonstrated technology and methodology for rapid environmental monitoring deployments.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[DTI](#) - Danish Technological Institute / Mikkel Labori Olsen
[ERATOSTHENES Centre of Excellence](#) / Dr. Michalis Mavrovouniotis and Stelios Neophytides

Industrial Domain: Space

IDS.4 | Enhancing natural disaster response through satellite-based earth observation and language models

Keywords: satellite imagery, natural disasters, earth observation, small language models, large language models

STATE-OF-THE-ART

Natural disasters such as wildfires, floods and landslides are escalating in occurrence and intensity due to climate crisis. Earth Observation (EO), and more specifically satellite-based remote sensing have been proven crucial sources of data for the post-event assessment providing insightful information to first responders and relevant stakeholders. Traditional EO processing, Machine Learning and Deep Learning are useful technologies for the processing of such data, but the process remains labour-intensive.

Recent advances in Small and Large Language Models (SLMs/LLMs) opens new pathways in EO data processing. Such models can interpret metadata, extract detected changes in satellite images, generate human-readable reports and respond to human queries. The embedment of SLMs/LLMs into EO workflows will enable responders to receive immediate and context-aware support upon the availability of satellite data and bridge the gaps for decision-making.

SCIENTIFIC CHALLENGES

- Accurate interpretation of complex landscapes (e.g., appearance of smoke, clouds or haze) affected by natural disasters from multiple satellite data sources.
- Fusing raw or analysis ready satellite data with ancillary data (e.g., land cover maps, crop type maps, topography, weather, etc.).
- Fine-tuning of a general-purpose SLM/LLM for EO-based natural disaster tasks.
- Trust and explainability of SLMs/LLMs.

RESEARCH ACTIVITIES

- Preparation of the needed datasets.
- Fine-tuning of an SLM/LLM for detection and analysis (e.g., generate a human-readable report) of natural disasters.
- Develop an interactive proof-of-concept system that enables natural language queries (e.g., show all burned areas of Cyprus for summer 2025) and automatic generation of summaries from EO data.
- Validate language outputs against expert-written reports and geospatial ground-truth data.
- Disseminate findings through scientific journals and high quality conferences.

EXPECTED RESULTS

- A fine-tuned SLM/LLM for natural disasters analysis through EO data.
- A demonstration platform showcasing the capabilities of the developed SLM/LLM.
- At least 1 journal publication and 1 conference publication.

POSSIBLE HOST ORGANISATIONS / SUPERVISORS

[ERATOSTHENES Centre of Excellence](#)- / Dr. Michalis Mavrovouniotis and Stelios Neophytides