CALL FOR PROPOSALS

EXCHANGE SCHEMES

OC2-2024-TES-02

ENFIELD: EUROPEAN LIGHTHOUSE TO MANIFEST TRUSTWORTHY AND GREEN AI

Co-funded by
the European Union

TABLE OF CONTENTS

INTRODUCTION

This is the catalogue of challenges available to the second out of four open calls for individual researchers exchange under the ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI) project, co-funded by the European Union. 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 specific scientific/technological challenges in artificial intelligence, contributing to ENFIELD network creation and expansion to European AI labs.

Pillar Green-AI

G-AI.1: Green Generative Language Models

Keywords: Generative AI, Large Language Models, Energy Consumption, Environmental Impact, Predictive Process Monitoring

STATE-OF-THE-ART

Advancements in large language models (LLMs) have revolutionized natural language understanding and generation, but their substantial computational demands raise environmental concerns. Current research focuses on enhancing LLM energy efficiency through techniques like model pruning, quantization, and distillation, alongside optimizing training processes with efficient hardware and renewable energy, as well as improving fine-tuning and prompting techniques, e.g., leveraging in-context learning (ICL). Edge computing is also being explored to distribute processing loads, reducing reliance on energy-intensive data centres. These efforts aim to create powerful yet environmentally sustainable LLMs. On the application side, using pre-trained LLMs with ICL could also reduce energy consumption by adapting to diverse tasks without costly retraining.

SCIENTIFIC CHALLENGES

The challenge focuses on making generative language models and their deployment greener by optimizing with regards to both their energy consumption and performance. Researchers will analyse the energy consumption of various LLMs to identify patterns and inefficiencies and investigate mitigation methods. Key scientific challenges include developing criteria for selecting energy-efficient LLMs for specific tasks without compromising performance, managing temporal credit assignment for delayed feedback, and aligning AI objectives with user goals to avoid suboptimal outcomes. The research may also explore techniques like model pruning, quantization, and distillation to reduce computational load. Barriers include the complexity of accurately measuring energy use in diverse environments, the scarcity of high-quality data, and ethical concerns like privacy and bias. Researchers will access existing LLMs, energy-efficient computing resources, and datasets to evaluate energy consumption and performance, enabling the development of innovative, sustainable methodologies for deploying LLMs for various tasks. One application of interest is predictive process monitoring (PPM), where LLMs can analyse patterns in event log data from business processes like order tracking.

RESEARCH ACTIVITIES

Energy consumption analysis: Conduct comprehensive analyses of energy consumption patterns across various LLMs to identify inefficiencies and areas for improvement; Energy mitigation strategies: Explore and implement techniques such as model pruning, quantization, and distillation to reduce energy consumption while preserving performance standards. Optimization of LLM selection criteria: Develop and refine criteria for selecting the most energy-efficient LLMs tailored to specific tasks, ensuring optimal performance without excessive energy use; Analysis of trade-offs with fine-tuning/ICL: Studying trade-offs between LLM performance and energy consumption in different fine-tuning and ICL setups for predictive process monitoring; Temporal credit assignment management: Investigate and address challenges related to temporal credit assignment for delayed feedback, enhancing the sustainability of LLM learning processes; AI objective alignment: Develop methodologies to align AI objectives with user goals, ensuring efficient and effective interactions between AI systems and human users, avoiding suboptimal outcomes; Ethical considerations: Seamlessly incorporate ethical considerations, such as privacy and bias, into the development and deployment of sustainable LLM methodologies to ensure responsible AI use.

EXPECTED RESULTS

The ENFIELD project aims to achieve significant scientific progress in the development of greener generative language models and understanding energy consumption and performance trade-offs in their applications. Expected outcomes include the establishment of energy-efficient methodologies for analysing and optimizing LLMs, novel techniques for mitigating energy consumption in LLMs, criteria for selecting energy-efficient models, insights on energy-efficient fine-tuning/ICL setups, and strategies for aligning AI objectives with user goals. The expected results from the exchange include publications in scientific conferences or journals. Ultimately, the project seeks to advance the state of the art in sustainable AI by providing actionable insights and methodologies for deploying greener generative language models in real-world applications.

POSSIBLE HOST ORGANISATIONS

- SINTEF Digital, Department of Sustainable Communication Technologies (Erik Johannes Husom and Sagar Sen)
- Telenor Research & Innovation (Mike Riess and Jeriek Van den Abeele)

Pillar Green-AI

G-AI.2 Green AI in the Edge-to-Cloud Continuum

Keywords: Hybrid Models, physics-informed machine learning, Bayesian modeling, causal models, capabilities and fundamental limits

STATE-OF-THE-ART

Green AI research is evolving to address sustainability concerns in the edge-cloud continuum, emphasizing energy-efficient protocols for AI tasks to minimize carbon emissions. The transition from cloud to edge processing necessitates addressing challenges in task distribution, coordination, and data transfer optimization while ensuring low-latency model distribution and adaptation. Understanding the ecological implications of diverse hardware systems through life cycle assessment (LCA) is becoming increasingly important. Hybrid AI models integrating symbolic and data-driven approaches offer promise for optimization, while continual learning technologies facilitate self-improvement of AI systems. Scientific challenges include developing hybrid models to reduce environmental footprint and leveraging knowledge-based systems to streamline training and inference efforts.

SCIENTIFIC CHALLENGES

Scientific challenges for green AI in the edge-cloud-continuum:

- Reducing the energy consumption and runtime of simulations via trained surrogate models
- Assessing the simulation costs for creating training data and the increase in energy efficiency by using surrogate models
- Assessing the trade-offs between energy, runtime, and accuracy in hybrid models
- Incorporating knowledge into AI systems via feature engineering, regularization, and architecture selection
- Quantifying the resulting energy savings (training/inference) when prior knowledge is incorporated
- Assessing the trade-offs between accuracy, requirements for training data, and energy costs in training and inference

RESEARCH ACTIVITIES

- Hybrid models for reduced environmental footprint in various fields of application: surrogate models for (certain parts of) finite element simulations, agent-based simulations, etc.
- Knowledge-based systems for reduced training and inference efforts
- Bayesian approaches to include prior knowledge
- Consideration of causal models for data generation during model selection/training

EXPECTED RESULTS

By collaborating with researchers, the project seeks to develop energy-efficient hybrid models and knowledge-based systems that minimize environmental impact. Expected outcomes include enhanced methods for reducing the energy consumption and runtime of simulations, optimizing data transfer and task distribution, and incorporating prior knowledge into AI systems to streamline training and inference efforts. These innovations will contribute to the creation of sustainable AI solutions, offering a balance between performance, accuracy, and energy efficiency, and ultimately driving progress in environmentally conscious AI development.

POSSIBLE HOST ORGANISATIONS

KNOW Center, Department of Methods & Algorithms for AI (Bernhard Geiger and Franz Rohrhofer)

Pillar Green-AI

G-AI.3 Green AI Metrics

Keywords: Green AI Metrics, Computational Cost, Energy Usage Monitoring, Energy Efficiency, Sustainability

STATE-OF-THE-ART

State of the art in monitoring Green AI metrics is focused on developing standardized, accurate metrics to measure the environmental impact of AI throughout its lifecycle. These metrics aim to evaluate AI architectures not only for performance accuracy but also for energy efficiency and reduced carbon emissions, including hardware manufacturing impacts. Challenges include standardizing these metrics, data scarcity for environmental impact assessment, and the need for collaboration across disciplines. Innovations in this field could lead to new measurement methods, tools, and principles for energy-efficient AI, providing a competitive edge while benefiting the environment. Significant work is also directed at estimating the computational efficiency, like floating-point operations (FLOPs), for various AI models, facilitating comparisons under fixed computational budgets, crucial for SMEs with limited resources. Furthermore, sectors like telecommunications are exploring dynamic management of network capacities using AI to reduce energy consumption. In industry, there's a movement towards integrating Green AI principles into system development to promote efficiency and robustness without relying solely on the latest hardware advances.

SCIENTIFIC CHALLENGES

One of the primary scientific challenges in monitoring Green AI metrics lies in the establishment and standardization of these metrics across varied AI system architectures. This includes not only the computation of the efficiency and accuracy of algorithms but also accounting for the environmental impact throughout the AI lifecycle, from hardware production to operational deployment.

Researchers must grapple with the dearth of universally accepted metrics and the limited availability of comprehensive data needed to assess the full environmental footprint of AI technologies. This challenge is compounded by the need for interdisciplinary collaboration to ensure that any developed metrics are both technically sound and environmentally meaningful. The goal is to create a suite of standardized Green AI metrics that balances performance with energy efficiency, guiding the design of AI systems that are both robust and sustainable. For a comprehensive proposal, researchers would require access to current AI models, energy consumption data, and cross-sectoral environmental impact assessments, alongside the tools for the simulation and evaluation of AI architectures under these new metrics.

RESEARCH ACTIVITIES

Multiple tools exist today and can be used to estimate in real-time the carbon emissions of training and using AI models. The research activities here should focus on filling the gaps by providing new tools in contexts that were not considered before. At the code level, existing tools are generally Python libraries that can be used when developing AI-based algorithms. There is a need for new real-time tools, programmed in different and more low-level languages such as C/C++. The conducted research will lead to the development of C/C++ tools/libraries dedicated to providing different metrics related to CO2 emissions and electricity usage when training and using AI models. Focus will be made on deep learning algorithms and industrial scenarios, involving robotics and machine vision. Contrary to LLM and models applied to cross-section data, there is a gap in measuring the emissions of time series algorithms. Particularly, most of the competitions that assess the performance of these algorithms (e.g., Makridakis competitions) just focus on prediction capabilities (sometimes in both, point forecasts and prediction intervals). Therefore, there is a need to include computational efficiency metrics in these competitions. We consider that metrics mixing forecasting power and efficiency (low computational cost - emissions) should be developed and tested, so that time series models could be compared and ranked in these two dimensions.

EXPECTED RESULTS

The research results will lead to the development of new tools and new benchmarks, which could be beneficial for anyone developing AI-based algorithms. The ENFIELD project aims to provide to the scientific community novel tools, metrics and guidance to develop greener algorithms in various use cases. Allowing AI developers to estimate CO2 emissions and electricity usage is a critical and essential step towards more sustainable AI.

POSSIBLE HOST ORGANISATIONS

- UCM, Economic Analysis Department & ICAE (Alfredo García Hiernaux)
- DTI, Robot Technology Center (Francois Picard)

Co-funded by the European Union

INOVA+ RESEARCH & TECHNOLOGY INNOVATION FOR GROWTH

Pillar Adaptative-AI

A-AI.1 Adaptive AI on the Edge – Hardware Aware Differentiable Architecture Search

Keywords: Hardware-Aware Architecture Search; Deep Learning; Neural Network; NAS; DARTS

STATE-OF-THE-ART

Deep neural networks (DNNs) are gaining popularity in a wide variety of embedded intelligent scenarios, such as computer vision, virtual reality, object detection and tracking, etc., offering impressive performance and enabling entirely new experiences on devices. Architecture design plays a crucial role for DNNs. Nevertheless, since the network design space is very large, the manual design of competitive DNNs requires enormous engineering efforts to determine the optimal network configuration, such as its depth and width. For this reason, neural architecture search (NAS), which aims to automate the design of high-quality DNNs, has recently flourished.

In the literature, studies on neural architecture search are mainly divided between reinforcement learning, evolution, and gradient-based or differentiable categories. However, the reinforcement learning and evolution-based NAS approaches suffer from substantial search overhead (several thousand GPU days), while the differentiable solution has significant search efficiency that reduces the search cost by several orders of magnitude. Due to its high search efficiency, the differentiable NAS scheme has recently emerged as the most dominant alternative in the NAS community.

SCIENTIFIC CHALLENGES

Despite significant progress, early differentiable NAS approaches such as DARTS are hardware indifferent. Specifically, they focus on finding architectures that are competitive in terms of accuracy, without considering other critical performance constraints such as latency, energy, and memory. Thus, they arrive at the architecture that offers promising accuracy for the target task, but at the cost of high computational complexity on the target hardware, which hinders the practical deployment of DNNs, especially on resource-constrained embedded platforms.

RESEARCH ACTIVITIES

In this project, we address these challenges by exploring existing solutions and proposing a hardware-aware DARTS framework that automatically searches for efficient DNNs on different hardware types, taking into account constraints related to computing, energy and memory capacities. Specifically, we will propose to learn a differentiable function to approximate hardware performance on a given hardware platform. As a result, this function provides hardware feedback that can be directly integrated into the DARTS flow, allowing the search for efficient and accurate architectures.

EXPECTED RESULTS

Based on the host institution's existing solutions in this topic, the research results of this project may be submitted to a rank A*/A conference or equivalent.

POSSIBLE HOST ORGANISATIONS

Telecom Paris / IMT, Department of Computer Science and Networks (Van-Tam Nguyen)

Pillar Adaptative-AI

A-AI.2 Adaptive AI on the Edge – Resources-constrained Training

Keywords: On-device training; task-adaptive sparse-update; gradient approximation; low-rank backpropagation

STATE-OF-THE-ART

On-device learning adapts pre-trained AI model to newly collected data after deployment. By training and adapting locally at the edge, the AI model can learn to improve its predictions and achieve continuous learning and user personalization. For example, adjusting a language model enables continuous learning from users' typing and writing; adapting a vision model enables recognition of new objects from a mobile camera. By bringing the training closer to the sensors, it allows, among other things, to protect the privacy of users when processing sensitive data (e.g., healthcare). However, on-device learning on tiny IoT devices is extremely difficult and fundamentally different from training in the cloud. Small IoT devices (e.g., microcontrollers) typically have a limited SRAM size, on the order of 256 KB. Such a small memory budget is barely sufficient for deep learning model inference, let alone training, which requires additional computation for backward and additional memory for intermediate activation. On the other hand, modern deep learning frameworks (e.g., PyTorch, TensorFlow) are typically designed for cloud servers and require a large memory footprint (>300 MB) even for small model training.

SCIENTIFIC CHALLENGES

The huge gap (>1000×) makes it impossible to run on edge devices with current frameworks and algorithms. Current deep learning systems like PyTorch, TensorFlow, etc. do not take into account the limited resources systems. Edge deep learning inference frameworks such as TVM, TF-Lite, NCNN, etc. offer reduced execution time, but do not support backpropagation. Although there are efficient and low-cost transfer learning algorithms, such as train to classifier final layer, bias update only, etc., the accuracy drop is significant and existing training systems cannot turn theoretical savings into measured savings. Furthermore, devices such as microcontrollers lack the operational system and runtime support needed by existing training frameworks. Therefore, we need to co-design the algorithm and system to enable tiny learning on the edge devices.

RESEARCH ACTIVITIES

- Task-adaptive sparse-update algorithms that dynamically selects the layer/channel based on a multi-objective criterion (the memory and the compute capabilities)
- HOSVD-based Gradient Approximator taking into account multi-objective criterion (the memory and the compute capabilities)
- Low-rank backpropagation (e.g., Walsh-Hadamard Transformation)

EXPECTED RESULTS

Based on the host institution's existing solutions in this topic, the research results of this project may be submitted to a rank A*/A conference or equivalent.

POSSIBLE HOST ORGANISATIONS

Telecom Paris / IMT, Department of Computer Science and Networks (Van-Tam Nguyen)

Pillar Adaptative-AI

A-AI.3. Adaptative Robustness and Trustworthiness

Keywords: Interpretable Representational Alignment, Disentanglement, Neurosymbolic Approaches Robustness, Conformal and cautious learning

STATE OF THE ART

Existing AI systems have difficulty adapting to dynamic environments. Adaptive AI systems are therefore needed in situations where rapid changes in the external environment or evolving corporate objectives demand an optimized response. State-of-the-art AI systems currently in use have significant shortcomings, notably lack of robustness in the face of disruption or of high uncertainty, low interpretability of outcomes, and so on. One of the main shortcomings of the de facto standard for deployed AI systems is that they assume that data and therefore the context in which they evolve is static, whereas it evolves. This requires the development of Adaptive AI systems that are versatile and robust to changes in the external environment as well as being interpretable and with strong reliability guarantees. This requires models capable modulating learned representations based on knowledge of the context or environment, but also minimizing the impact on robustness of an evolving and uncertain context on prediction accuracy, particularly when sequential predictions are involved.

SCIENTIFIC CHALLENGES

Adaptive and Interpretable Representational Alignment, Neurosymbolic learning
How to integrate knowledge in Adaptive AI systems by aligning representations from foundation models (one or more modalities) and from knowledge graphs (i.e. using knowledge graph embedding techniques), by conditioning the representations on knowledge structure through disentanglement.
How to exploit these representations for enhanced interpretability, robustness and trust?
Cautious models for robust Adaptive AI
How to enforce guarantees (either statistically or deterministically) on the accuracy of the predictions within a given operational domain, particularly when predictions are sequential and where errors can accumulate, potentially leading to disastrous consequences?

RESEARCH ACTIVITIES

Activities to address the challenges include:

- Surveying existing techniques and identifying gaps for methodological contributions
- Identifying appropriate use-cases from the verticals (particularly manufacturing and healthcare) or related to the objectives and main challenges addressed by the verticals in collaboration with the host institution.
- Developing the methodological innovation and evaluating it against standard benchmarks as well as dataset from or related to the verticals.
- Writing a high-quality publication in top venues (ICLR, NeurIPS, CVPR, ICCV, ECCV, *ACL, AAAI, IJCAI, interdisciplinary Q1 journals for use-case applications, e.g., Journal of Biomedical Informatics for health)

EXPECTED RESULTS

The ENFIELD project aims to foster contributions on data-driven and knowledge-based methods for enabling adaptations in learning. For this call, the focus is twofold: on interpretable representational alignment through neurosymbolic learning and on Cautious Classification systems in dynamic settings. We hope to establish new but lasting collaborations on these topics between experts outside the consortium and partners inside the consortium, where the mobility is a first step and a first contribution of the hopefully fruitful collaboration.

POSSIBLE HOST ORGANISATIONS

- IMT (priority targets IMT Mines Alès, Télécom Paris, but any IMT site possible so long as a suitable PI can be identified)
- TU/e, Department of Mathematics and Computer Science

Co-funded by the European Union

Pillar Adaptative-AI

A-AI.4 - Brain-inspired AI - Developing intelligent systems that are effective and efficient continual learners

Keywords: Continual learning, multi-modal learning, robustness, foundation models, Vision-language models (VLMs)

STATE OF THE ART

Deep neural networks (DNNs) are still not on par with human brain intelligence. Most of the existing Continual Learning methods are computationally expensive and ineffective. They fail to mimic the intricacies of the learning mechanisms and the interactions of multiple memory systems in the human brain. More and more recent methods have drawn inspiration from the brain e.g. experience replay, synaptic consolidation and multiple memory systems which have shown promise and make a strong case for further work in this promising direction. More sparse representation and multi-modalities, which is how brains learn are also starting to become prominent in AI systems.

SCIENTIFIC CHALLENGES

- Humans excel at continual learning, seamlessly integrating new information without forgetting previous knowledge, while DNNs suffer from catastrophic forgetting. Additionally, humans demonstrate robust learning and adaptability, whereas DNNs are vulnerable to adversarial attacks, easily being misled by minor input alterations. These limitations highlight the need for more advanced, brain-inspired approaches in AI research.
- Better generalization and lifelong learning- Identifying the gaps between Humans and Existing AI presents a unique opportunity to revisit the design of DNNs to enhance their continual learning capabilities and generalization.
- **Efficiency through the lens of sparsity** - The overarching objective is to lead the charge in developing cutting-edge sparsification techniques meticulously crafted to enable efficient continual learning on resource constrained devices.
- **Leverage Multiple Modalities** - A salient feature of the brain that may play a critical role in enhancing its lifelong learning capabilities is that it processes and integrates information from multiple modalities. Hence, we aim to combine information from different modalities which allows the models to develop a more comprehensive understanding of the environment.

RESEARCH ACTIVITIES

- **Intelligent and Efficient Continual Learning** – Explore brain-inspired concepts to improve continual learning capabilities. Further take inspiration from sparse coding in the brain to develop more sparse and efficient algorithms.
- **Multi-modal learning / Foundation models**– Explore different modalities in deep learning and approaches to leverage the complementary information in each modality to learn a more holistic and robust representation of the objects; Explore ways to leverage foundation models (LLMS or VLMS) to improve generalization and reasoning in vision encoders.

EXPECTED RESULTS

The ENFIELD project aims to bridge the gap between the capabilities of humans and existing AI. By leveraging insights from our enhanced understanding of the brain, it aims to design the next generation of brain inspired models that enable efficient and effective continual learning in deep neural networks. The goal is to designs AI models that can be deployed in our dynamic environment and meet the ever-changing industrial requirements.

POSSIBLE HOST ORGANISATIONS

TU/e, Department of Mathematics and Computer Science (Bahram Zonooz)

Co-funded by the European Union

INOVA+ RESEARCH & TECHNOLOGY INNOVATION FOR GROWTH

Pillar Adaptative-AI

A-AI.5 - Brain-inspired AI – developing intelligent systems that are better in generalization and robustness

Keywords: RobustAI, adversarial robustness, generalization, language models, vision-language-models

STATE OF THE ART

Deep neural networks (DNNs) still fall short of human brain intelligence, particularly in terms of robust learning and adaptability. While humans can seamlessly adjust to new information, DNNs are susceptible to adversarial attacks, where slight modifications to inputs can lead to significant misclassifications. Techniques such as adversarial training aim to mitigate this vulnerability but face challenges like balancing standard and adversarial accuracy and dealing with robust overfitting. Despite advancements, achieving robust AI that parallels human cognitive resilience remains a complex and ongoing endeavour.

SCIENTIFIC CHALLENGES

Humans exhibit robust learning and adaptability, whereas DNNs are vulnerable to adversarial attacks, being easily misled by minor input alterations. These limitations underscore the need for more advanced, brain-inspired approaches in AI research.

To enhance robustness and generalization - against both natural and adversarial perturbations. Improving and proposing novel training schemes to address the trade-off between standard and adversarial accuracy and tackling robust overfitting. The robustness against different perturbations should be analysed and evaluated across vision-only models, language models, and multi-modal models, ensuring comprehensive and resilient AI systems capable of performing reliably in diverse and challenging scenarios.

RESEARCH ACTIVITIES

- Conduct a comprehensive literature review on natural and adversarial robustness - evaluate different adversarial training techniques and other robustness methods.
- Assess the robustness of vision, language, and multi-modal models against natural and adversarial perturbations.
- Propose a new methodology based on insights from the evaluations, aiming to surpass current state-of-the-art techniques.
- Run evaluation on an application (from the verticals)
- Document findings and submit a paper to a reputable conference or journal

EXPECTED RESULTS

The ENFIELD project aims to bridge the gap between the capabilities of humans and existing AI. By leveraging insights from our enhanced understanding of the brain, it seeks to design the next generation of brain-inspired models that enhance robustness in deep neural networks. The goal is to develop AI models resilient to both natural and adversarial perturbations, enabling reliable deployment in dynamic environments and meeting ever-changing industrial requirements.

POSSIBLE HOST ORGANISATIONS

TU/e, Department of Mathematics and Computer Science (Bahram Zonooz)

Pillar Human-Centric-AI

HC-AI.1 Evolving Symbolic Models for Decision-Making

Keywords: Symbolic AI; Reinforcement learning; Learning; Data-driven; Evolving.

STATE-OF-THE-ART

Neuro-symbolic learning uses context-free grammar (from automata theory) as a symbolic representation and learns from an oracle (i.e., an artificial neural network trained with reinforcement learning) in a supervised learning setting (imitation learning)1. As far as we know, the only publication that does not use an oracle is2, which conducts a program architecture search on top of a continuous relaxation of the architecture space defined by programming language grammar rules.

Another research direction is iterative machine learning, where humans are part of the learning process and can tune the hyperparameters of a meta-heuristic optimizer3. Genetic programming for symbolic regression is also an alternative for learning symbolic models from data. Still, it uses a tree-based representation for the knowledge that can be ineffective for complex symbolic structures. An interesting work is4, where evolutionary search with a list of 65 basic mathematical operations is used to discover ML algorithms from scratch with minimal human intervention automatically.

SCIENTIFIC CHALLENGES

Grow symbolic models from data based on the interaction between the AI-decision system and the environment, where reinforcement learning can be used for constructing the model. The human can define a template for the symbolic model and/or participate in the learning phase (e.g., change hyperparameters, modify intermediary solutions).

RESEARCH ACTIVITIES

- Study different symbolic representations for control/decision problems (e.g., from automata theory). This representation might be domain-specific, which means having a domain-specific language for each use case.
- Development of symbolic search methods based on well-established search-based algorithms such as simulated annealing or Monte Carlo Tree search.
- Application of the developed approach in verticals (namely in Energy).

EXPECTED RESULTS

The expected scientific progress includes the development of a new method for symbolic AI that is capable of learning from data within a reinforcement learning framework. The method can be used to augment existing expert systems in different domains, e.g., use the existing expert system as a template or starting point for learning or find new structures and symbolic representations for those systems. It should offer higher interpretability to humans since they are part of the learning process in three possible stages: (1) design of the model's template/structure, (2) modify or improve solutions during the learning phase (iterative learning), and (3) analyse and modify the final solution.

The beneficiary is expected to have one peer-reviewed publication (preferably in a top-tier journal or conference) and the code published in open-source (GitHub).

POSSIBLE HOST ORGANISATIONS

INESC TEC, Center for Power and Energy Systems (Ricardo Bessa)

Pillar Human-Centric-AI

HC-AI.2 User Perspectives on Explainable AI

Keywords: Explainable AI, Human-centric AI, Human-AI Interaction, Trustworthiness, Behavioural Science

STATE-OF-THE-ART

Real-world adoption of AI necessitates trust from users to ensure its effective integration in technical systems. Current research highlights the importance of explainability in AI to improve human-AI collaboration, enabling users to understand and trust AI decisions. Studies focus on techniques such as confidence scoring and local explanations for predictions, which may help with calibrating trust where appropriate. The effectiveness of different techniques can vary greatly depending on the level of detail and presentation given to the users. The skill level of users is also a factor to be considered in the analyses, on which there is not much evidence in the literature yet.

SCIENTIFIC CHALLENGES

This challenge addresses the AI explainability needs of technical systems users. Researchers will investigate user preferences for AI solution design through interviews and surveys, using realistic scenarios and mock-ups. The study aims to determine which aspects of AI explainability can enhance trust and whether the provided information is useful for decision-making. Key scientific challenges include understanding the balance between transparency and complexity, identifying the most effective forms of explanations, and measuring the impact of these explanations on user trust and actions in different decision settings with varying impact of the decisions. The research will involve mapping current AI perceptions and practices of technical system users, followed by user testing of various explainability solutions. Barriers include the subjective nature of trust, variability in user preferences, and the challenge of creating universally effective explainability features.

RESEARCH ACTIVITIES

Research activities will begin with mapping the baseline practices and perceptions of selected technical staff through interviews to understand how data-driven their current operations are and their views on and attitudes towards AI. Based on the input from the interviews, a survey targeted at a broader audience of technical system users will be conducted. As part of the survey, participants will be exposed to different AI explainability solutions using mock-ups. The study will use A/B testing to compare the effectiveness of various explainability features. Key activities include developing and refining the mock-ups, conducting interviews and surveys, analysing the data to identify trends and preferences, and synthesising these insights to inform AI solution design. Interdisciplinary collaboration will be essential to address the human-centric and technical aspects of explainable AI. Feedback loops will ensure that the study evolves based on participant input and emerging findings.

EXPECTED RESULTS

The ENFIELD project aims to gain valuable insights into the design of human-centric explainable AI solutions that will increase their usability and real-world adoption. Expected outcomes include a detailed understanding of technical users' preferences for AI explainability features, the identification of key factors that influence user trust in AI-assisted decisions, and actionable recommendations for AI solution design. Scientific results will include quantitative and/or qualitative data on the effectiveness of different explainability methods, their impact on user trust and decision-making, and guidelines for integrating explainability into AI systems. Evaluation measures will include assessments of user trust, satisfaction, and the perceived usefulness of explainability features. Ultimately, the project seeks to advance the adoption of AI by providing practical insights into designing user-friendly explainable AI solutions.

POSSIBLE HOST ORGANISATIONS

Telenor Research & Innovation (Jeriek Van den Abeele and Eleonora Freddi)

HC-AI.3 - Child-Centric AI for Personalized and Safe Educational Voice Assistants

Keywords: speech technology, personalized, voice assistant, child speech, speech synthesis

STATE-OF-THE-ART

Educational voice assistants play an important role in children's learning experiences. However, ensuring both personalization and safety in these applications remains a challenge.

SCIENTIFIC CHALLENGES

- Develop AI models for personalized educational voice assistants that provide to individual learning needs.
- Create safety mechanisms and monitoring features to protect children while using educational voice assistants.
- Address privacy and ethical concerns in the context of child-specific voice data.

RESEARCH ACTIVITIES

- Develop AI models specialized in personalizing educational voice assistants for children.
- Investigate methods for implementing safety features, parental controls, and monitoring tools in educational voice assistant applications.
- Implement robust privacy measures to safeguard child-specific voice data.

EXPECTED RESULTS

1 scientific publication (conference paper)

POSSIBLE HOST ORGANISATIONS

BME, Department of Telecommunications and Artificial Intelligence (Géza Németh)

Pillar Human-Centric-AI

HC-AI.4 - Novel Explainable AI Methods for Decision Making

Keywords: Explainability; Spatio-temporal models; Decision making; Healthcare; AAL

STATE-OF-THE-ART

Models such as graph neural networks or multi-modal Transformers may be powerful tools for modelling different dependencies in spatio-temporal contextual relationships. One challenging area of application is Human Action Recognition (HAR) from video sequences, especially when performed in real-world environments, for example in Healthcare, Ambient Assistive Living (AAL) or Manufacturing. Different DNN architectures, such as GNN, TCN, or Spatial Temporal GCN, have been proposed for solving the HAR problem.

Spatio-temporal transformers have emerged as a promising approach for modeling spatio-temporal relationships, for example Video action Transformer Network, ConvTransformer Network, or Spatio-Temporal Attention Network.

However, their black box nature limits their interpretability for trustworthy decision making. The interactions between the space and time dimensions add a layer of complexity, making it difficult to trace how inputs influence the outputs. There are still few approaches that try to explain how such networks arrived at a decision or, most important, why they failed to predict the correct result.

SCIENTIFIC CHALLENGES

- The main challenge in explainability for spatio-temporal transformers lies in their complexity that may require different explainability techniques.
- Understand user's requirements for explainability and interpretability of spatio-temporal models.
- Combine multiple explainability techniques, for example using attention visualizations together with feature attribution methods, to offer a deeper insight into how different data types interact and influence the final output.
- Find new explainability methods for spatio-temporal transformer architectures without increasing their inner complexity.
- Design a high-level representation structure to capture the spatial and temporal dimension of human action, as a basis for linking DNN explanations to human decision making.

RESEARCH ACTIVITIES

- Design and implement explainable algorithms for multi-modal transformers.
- Design new XAI methods to capture the flow of information and dynamics in spatio-temporal structures.
- Explore different approaches to link explanations to symbolic structures.
- Evaluate from a qualitative and a quantitative point of view the quality of explanations.

EXPECTED RESULTS

The expected scientific progress includes the development of high-level XAI models for decision-making applications based on spatio-temporal models. These models may be used in different applications, for example in Healthcare, AAL or Manufacturing (Human-AI collaboration). The beneficiary is expected to develop demos for these models in one or more of the above-mentioned application domains and to produce one peer-reviewed publication (preferably in a top-tier journal or conference).

POSSIBLE HOST ORGANISATIONS

- UPB, Artificial Intelligence and Multi-Agent Systems Laboratory (Adina Magda Florea)
- TU/e, Department of Industrial Engineering and Innovation Sciences (Isel Grau Garcia)
- INESC TEC, Center for Power and Energy Systems (Ricardo Bessa)

Pillar Trustworthy -AI

T-AI.1 Secure Voice Biometrics with Fake Voice Detection

Keywords: Voice spoofing, Biometric security, Speech signal processing, Robust authentication, Acoustic analysis

STATE-OF-THE-ART

Current voice biometric systems are at the central of biometric authentication, but they face increasing concerns related to data privacy and security. The rise of fake voice generation technologies presents a significant challenge to the integrity of voice biometrics. State-of-the-art solutions in this field are actively addressing the need to develop robust defences against not only traditional security risks but also the voice spoofing and deepfake technologies. As the use of voice biometrics continues to expand in applications like access control, financial transactions, and identity verification, it is essential to address these scientific challenges and opportunities to ensure the trustworthiness and reliability of voice-based authentication methods.

SCIENTIFIC CHALLENGES

- Developing AI models to enhance the security and trustworthiness of voice biometrics is a complicated task. This requires the creation of superior algorithm capable of distinguishing real from synthetic voices. This challenge requires the combination of cutting-edge deep learning techniques with voice recognition to continuously adapt and secure against emerging threats. This involves dealing with various accent and language variations, background noise, and voice quality issues.
- Protecting sensitive voice data is another crucial aspect. This challenge involves developing mechanisms that safeguard stored and transmitted voice samples. It needs a deep understanding of data encryption and secure communication protocols designed to voice biometrics.
- Addressing voice spoofing is important because it directly impacts the reliability and security of voice biometric systems. As voice authentication becomes more common in various sectors, including finance, healthcare, and access control, the threat of voice spoofing presents a significant risk. Developing robust anti-spoofing techniques is necessary to ensure the trustworthiness and integrity of voice-based security measures, maintaining user confidence in the technology and safeguarding sensitive information against fraudulent activities. This challenge needs advanced signal processing, machine learning, and behavioural analysis methods.
- Advancing techniques for the detection of fake voice samples requires exploring speech characteristics and signal analysis. This challenge requires not only identifying synthetic voice attributes but also understanding how these attributes differ from natural human speech. This challenge needs deep learning, feature engineering, and acoustic analysis to design more accurate and reliable fake voice detection methods.

RESEARCH ACTIVITIES

- Develop AI models for secure voice biometrics, integrating encryption, privacy-preserving methods, and fake voice detection.
- Investigate methods for detecting and preventing voice spoofing, as well as the generation of fake voice samples.
- Perform accurate testing to ensure the system's trustworthiness, reliability, and fake voice detection capabilities.

EXPECTED RESULTS

Expected results involve the investigation of novel methods to detect and prevent voice spoofing, ensuring the system's strength against manipulated voice samples. Moreover, accurate testing will be conducted to verify the system's trustworthiness, reliability, and its capabilities in detecting fake voices.

POSSIBLE HOST ORGANISATIONS

BME, Department of Telecommunications and Artificial Intelligence (Géza Németh)

Pillar Trustworthy -AI

T-AI.2 Non-linear models for multivariate time series (vertical federated learning)

Keywords: Vertical Federated Learning, Data privacy, Interpretability, Kolmogorov-Arnold Networks, Fusion models

STATE-OF-THE-ART

Distributed time series data can improve forecasting accuracy but requires privacy measures. Current literature on data privacy focuses on horizontal federated learning (FL), where agents share features across instances. Vertical FL, where agents observe different features in the same instances, poses challenges due to indeterminate coefficients. Methods like adding random noise or using cryptographic techniques are available but computationally intensive with numerous time series. As AI regulation advances, interpretable solutions also become crucial.

SCIENTIFIC CHALLENGES

Many successful models rely on neural networks, which are often black-box models. Recent advances show that Kolmogorov-Arnold Networks (KANs), an interpretable neural network designed for complex function fitting, can match or exceed the accuracy of larger MLPs with smaller architectures. In this challenge, we propose exploring KANs and fusion architectures to protect privacy, aiming to improve both model accuracy and data privacy in AI systems while avoiding heavy encryption protocols.

RESEARCH ACTIVITIES

The research activities include:

- Exploring literature on privacy-preserving federated learning (FL);
- Investigating vertical FL with fusion models;
- Understanding Kolmogorov-Arnold Networks (KANs);
- Assessing the value of combining data from multiple sources (local and collaborative KANs) with real datasets from the energy sector (e.g., renewable energy production or load);
- Integrating KANs with fusion models to enhance data privacy and model performance;
- Disseminating the results from the previous points with a scientific publication on a journal or international conference.

EXPECTED RESULTS

The ENFIELD project aims to advance trustworthy, secure, and distributed AI systems. Researchers will develop privacy-preserving federated learning approaches based on interpretable neural networks. Expected outcomes include at least one scientific publication and a lasting collaboration between the visiting researcher and the hosting institution, extending beyond the research visit.

POSSIBLE HOST ORGANISATIONS

INESC TEC, Center for Power and Energy Systems (Carla Gonçalves)

Pillar Trustworthy -AI

T-AI.3 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

- NTNU, Department of Information Security and Communication Technology (Georgios Spathoulas and Georgios Kavallieratos)
- Telenor Research & Innovation (Jeriek Van den Abeele and Claudia Battistin)

T-AI.4 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, the 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 approached 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

- NTNU, Department of Information Security and Communication Technology (Georgios Spathoulas and Georgios Kavallieratos)
- TUC, Distributed and Self-organizing Systems (Sebastian Heil)

Pillar Trustworthy -AI

T-AI.5 - AI in Distributed systems

Keywords: Electric vehicles; Edge intelligence; Optimization; Renewable energy; Microgrid.

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 (AIaaS) 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 AIaaS 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 AIaaS 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

- NTNU, Department of Information Security and Communication Technology (Georgios Spathoulas)
- TUC, Distributed and Self-organizing Systems (Sebastian Heil)

Vertical Healthcare

VH.1 ICU readmission analysis and support

Keywords: ICU readmission risk, statistical correlation

STATE-OF-THE-ART

The prognosis for a patient that is readmitted in an Intensive Care Unit (ICU) is typically quite pessimistic, but hospitals do have limited capacity and need to prioritize events. Studies until now have identified some significant risk factors, such as patient condition before ICU admission, days already spent in the ICU, and specific medical/physical test scores. Also, there seems to be a higher probability of readmission for specific diseases. To that end, it is beneficial for patients, medical providers, and caregivers if high-risk of ICU readmission can be detected and considered as a critical factor before patient discharge.

SCIENTIFIC CHALLENGES

Studies until now focus on explainable Machine Learning techniques to analyse and interpret the vast amount of data, since the direct application of Artificial Intelligence to healthcare involves several ethical concerns, such as unfair algorithmic bias. This is a problem, as predictor models are treated as "black boxes", without understanding the internal performance and being unable to explain how a certain prediction is reached. Researchers have tried to use several deep learning methods to support physicians in decisions regarding patient discharge, but they usually focus on general ICU discharges with no regard to the specific diseases, conditions, and patient context. The combination of appropriate models and time-series analysis, along with state-of-the-art mathematical tools for correlation detection and open data adaptation can lead to interesting results.

RESEARCH ACTIVITIES

Research on this topic will need to start with interdisciplinary work on theoretical aspects and potential significant factors that can affect a specific type of ICU patient. We expect a specific survey on state-of-the-art, as well as search for relevant high-quality data. The researchers will need to be able to come up with a proposal for a potentially interesting novel approach, develop the decision support module and test it with real data. The applicants will need to have a good understanding of AI, in order to find the right models and techniques for the various parts of the problem and not just fit the problem to a specific type of model. It is imperative that the work and results are published in the end, so we expect one or two submissions to appropriate conferences and/or journals.

EXPECTED RESULTS

ENFIELD healthcare vertical will use the data gathered and the novel scientific results of all open calls to enhance own work on the subject, enrich the healthcare AI roadmap and propose the best practices and policies for such work in the near future. During the process, ICCS will promote knowledge exchange between ENFIELD and the involved institution(s) and support the effort for publication. ICCS is also interested in a potential follow-up of the work in a later stage, ensuring impact and serving the lighthouse paradigm. Furthermore, the results will be used as a proof of concept for potential collaborations with relevant healthcare stakeholders, such as hospitals, health ministries, and relevant policy makers.

POSSIBLE HOST ORGANISATIONS

NTUA, Computer Networks Laboratory (Ioanna Roussaki)

Co-funded by the European Union

INOVA+ RESEARCH & TECHNOLOGY INNOVATION FOR GROWTH

Vertical Energy

VE.1 Systemic and Local explainer of the system dynamic behavior

Keywords: Interpretable AI; Power System Dynamic Behaviour; Human Domain Knowledge.

STATE-OF-THE-ART

In the future power systems, where synchronous inertia will be lacking, and converters with different types of controls will be abundant, understanding and interpreting the system's dynamic behaviour will become increasingly more complex. Therefore, the task of area/zone operators in identifying weak links and optimal control variables will get trickier, especially when discussing phenomena of local nature. Typically, the operator conducts offline studies using dynamic simulation to extract practical insights and heuristics that are then used to improve system operation. Given the advances in explainable and interpretable AI and power system modelling, an opportunity arises for this task to be performed using an AI methodology capable of generating small explanations by leveraging an existing power system environment.

SCIENTIFIC CHALLENGES

The primary scientific challenge in this use case is to develop a methodology that can generate a concise set of easily understandable rules by utilising a high-fidelity system model. Given the computational complexity of power system dynamic simulation, the AI methodology should be lightweight, or an alternative approach for time-domain simulation should be suggested. In formulating rules, it is crucial to prioritise low mathematical complexity and incorporate system variables of both global and local nature (e.g., System Inertia and Short-circuit Ratio of Generator). The choice of variables for the AI methodology is still open for discussion. The selected variables should be able to explain the local dynamic behaviours of the system. Developing a system environment where local instability phenomena (e.g., loss of synchronism of specific converters or generators, tripping of specific protections, etc.) can be present and evaluated is also an important part of the challenge.

RESEARCH ACTIVITIES

- Development of a methodology capable of creating a small set of interpretable rules by leveraging real data on system events or a high-fidelity system model.
- Develop a high-fidelity power system model that can be used as an environment for the desired methodology.

EXPECTED RESULTS

The expected scientific progress includes the development of a new AI system capable of explaining to humans the dynamics of electric power systems. The beneficiary is expected to have one peer-reviewed publication (preferably in a top-tier journal or conference) and the code published in open-source (GitHub).

POSSIBLE HOST ORGANISATIONS

INESC TEC, Center for Power and Energy Systems (Ricardo Bessa)

Vertical Manufacturing

VM.1 Context-agnostic human detection in robotic cell

Keywords: Condition-based maintenance, tool machines, Health management, time series

STATE-OF-THE-ART

Collaborative robots are gaining an increasingly prominent role for what concerns the assistance to manufacturing operators, since this equipment is potentially capable of combining the productivity of robots with the flexibility of humans. Anyway, the occupancy of the same workspace by cobots and operators can lead, in case of impact, either to safety issues or to productivity losses. Thus, recent works1 are targeting the integration of Computer Vision (CV) functionalities, integrated by Digital Twins for virtual commissioning or for the generation of synthetic data needed for training/fine-tuning of CV algorithms. While these visual safety controllers have shown promise in controlled training environments, their performance can degrade significantly in real-world scenarios due to differences between the training data and real-world conditions (known as domain shift). To address this challenge, developing a domain-agnostic generalized visual controller is essential to ensure safety in human-robot collaboration.

SCIENTIFIC CHALLENGES

The hosting institution will make available an industrial setup close to the one depicted in the aforementioned work1, to allow the winner to start from an operational environment embodying the base already available in the literature. Besides of this, the proposal is supposed to address the following objectives:

- Development of a robust framework for human and robot detection and interaction recognition.
- Bridging the domain gap between training and inference time.
- Enhancing safety in human-robot collaboration for industrial applications.
- Improvement of performances from existing sources in literature.

RESEARCH ACTIVITIES

The candidate is expected to have a basic knowledge of robotics and a good knowledge of computer vision. In addition, it is expected for the candidate to be able to read and understand scientific papers in robotics and computer vision. Moreover, the following steps should be done:

- Investigating existing approaches and proposing a novel framework for visual human-robot interaction (HRI) detection and classification.
- Analysing the proposed framework's robustness against domain shift.
- Design and implement algorithms that improve and reduce the effect of the visual domain gap and create a visual domain agnostic model.
- Evaluating the effectiveness of the proposed method in different scenarios.

EXPECTED RESULTS

The expected results for this work should enhance human-robot interaction (HRI) safety by achieving visual distribution agnosticism during inference. By doing so, the visually automated HRI models should exhibit minimal performance degradation when deployed in new environments. This means the model can handle the same task even if the objects and robot's visual appearance differ from the

training data. The beneficiary is expected to develop a demo for such a task and to produce at least one peer-reviewed publication (preferably in a top-tier journal).

POSSIBLE HOST ORGANISATIONS

POLIMI, Department of Management, Economics and Industrial Engineering (Walter Quadrini)

Vertical Manufacturing

VM.2 Self-X Integration in manufacturing domain

Keywords: Manufacturing; self-X; autonomous computing; MAPE-K; self-CHOP

STATE-OF-THE-ART

The recent advancements in data production and analysis allowed several Machine Learning techniques to be exploited in manufacturing domain. One of the limits of these approaches, however, sits in the lack of resilience of these techniques towards unpredicted and unpredictable events, which drift the performances of these algorithms far from the region they were trained/designed for (e.g., a new production recipe is introduced, and the algorithms are not able to classify the system status). One of the most experienced solutions to these issues is the iteration of the training phases of algorithms, but this mitigation presents either high resource consumption or can lead to catastrophic interferences, which constitute a severe risk for the performances and for addressing responsibilities. In recent years, the practitioners' community resumed however frameworks and requirements from the Control domain (namely MAPE-K and Self-X) to grant the controlled system, the capability to self-adapt to unpredictable events.

SCIENTIFIC CHALLENGES

Literature in terms of Self-X functionalities and MAPE-K framework is small in number and varied, so the first challenge is represented by the need of clarifying definitions and domains of application. About these domains, the traditional application areas of Self-X functionalities have been localized into control loop-level modules, but the growing interest of scientific community in their application to higher level of the Computer Integrate Manufacturing structure needs some experimental evidences.

RESEARCH ACTIVITIES

Starting from a defined and centralised software architecture, the proposed solution is supposed to be able to make AI pipelines able to deal with and implement self-X capabilities.

The solution is supposed to be tailored onto a lab-scale production environment and to deal with non-PLC signals (e.g., energy consumption) clustering the production in new defined classes.

Alternatively, real-like industrial datasets are also available.

EXPECTED RESULTS

- At least one journal scientific publication (Scopus-indexed).
- Development of software modules implementing self-X functionalities.
- Experimentation in laboratory environment and/or from existing datasets.

POSSIBLE HOST ORGANISATIONS

POLIMI, Department of Management, Economics and Industrial Engineering (Walter Quadrini)

Co-funded by the European Union

INOVA+ RESEARCH & TECHNOLOGY INNOVATION FOR GROWTH