ECCO

European Cybersecurity COmmunity

AI supporting Cyber Risk Management and Resilience of Critical Infrastructures

30th September 2024

AI Supporting Cyber Risks And Resilience Of Critical Infrastructures

Insights From CS-AWARE NEXT, DYNABIC, & AI4CYBER

CS-AWARE CYBERSECURITY NEXT   AI4CYBER   DYNABIC

Join us **30 September 2024**

16:00 - 17:30 CEST

Prof. Juha Röning

CS-AWARE NEXT Project Coordinator, University of Oulo

Erkuden Rios

AI4CYBER and DYNABIC Project Manager, TECNALIA

Victor Muntés-Mulero

Exploitation and Innovation Manager at DYNABIC, Beawre

In collaboration with HORIZON RESULTS BOOSTER   European Commission   ECCC

Agenda

16:00 – 17:00 – AI Solutions supporting Cyber Risk Management

Design-Science and AI harnessed for improving ´Cybersecurity on infrastructure ecosystems
Prof. Juha Röning, Professor at the University of Oulu, CS-AWARE NEXT Coordinator
- Q&A

Smart Risk Management for Business Continuity in Critical Infrastructure
Victor Muntés-Mulero, Co-founder and CEO at Beawre, Exploitation and Innovation Manager at DYNABIC
- Q&A

Trustworthy AI for cybersecurity solutions
Erkuden Rios, AI4CYBER and DYNABIC Project Manager, TECNALIA
- Q&A

17:00 – 17:30 - Roundtable with speakers

Chair – Nicholas Ferguson, Trust-IT & ECCO

Funded by the European Union

Housekeeping

- The webinar is being recorded. A link to the full recordings will be shared with participants afterwards
- You’re welcome to ask questions! Please use Q&A panel to ask your questions: we will activate your microphone.
- You can also raise your hand during the dedicated Q&A time
- Roundtable at the end for further questions!

Funded by the European Union

Design-Science and AI Harnessed for Cybersecurity Infrastructure Ecosystem

Juha Röning
BISG
University of Oulu
Finland
Juha.Roning@oulu.fi

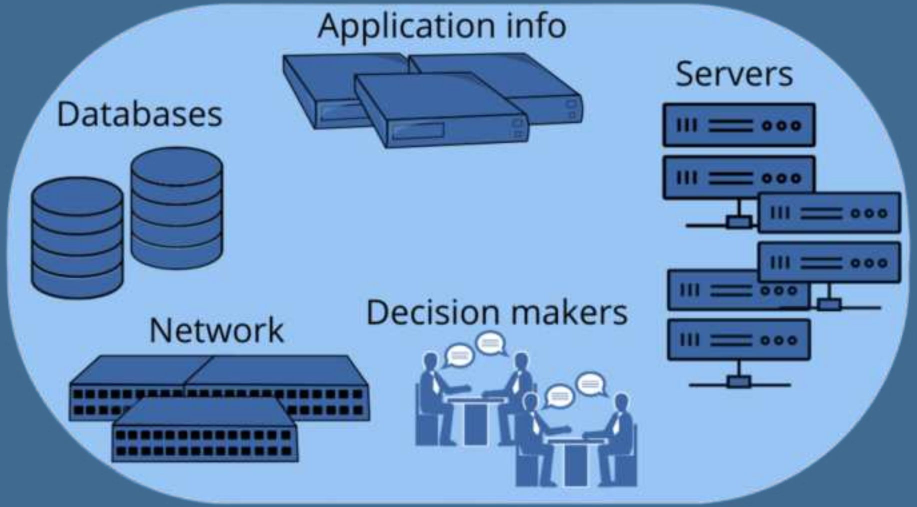
CS-AWARE CYBERSECURITY NEXT

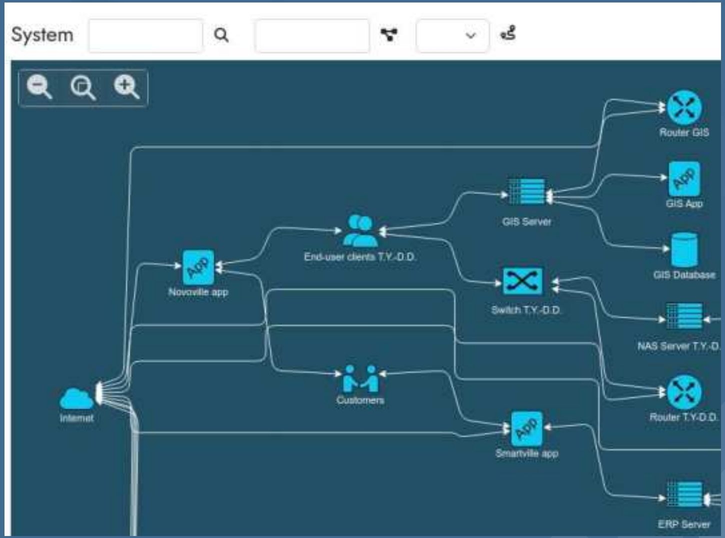
The CS-AWARE H2020 project

- A socio-technical systems based approach to cybersecurity management in organisations based on awareness
  - Humans are a central factor in any complex system
  - CS-AWARE enables humans to be part of the solution in solving cybersecurity issues

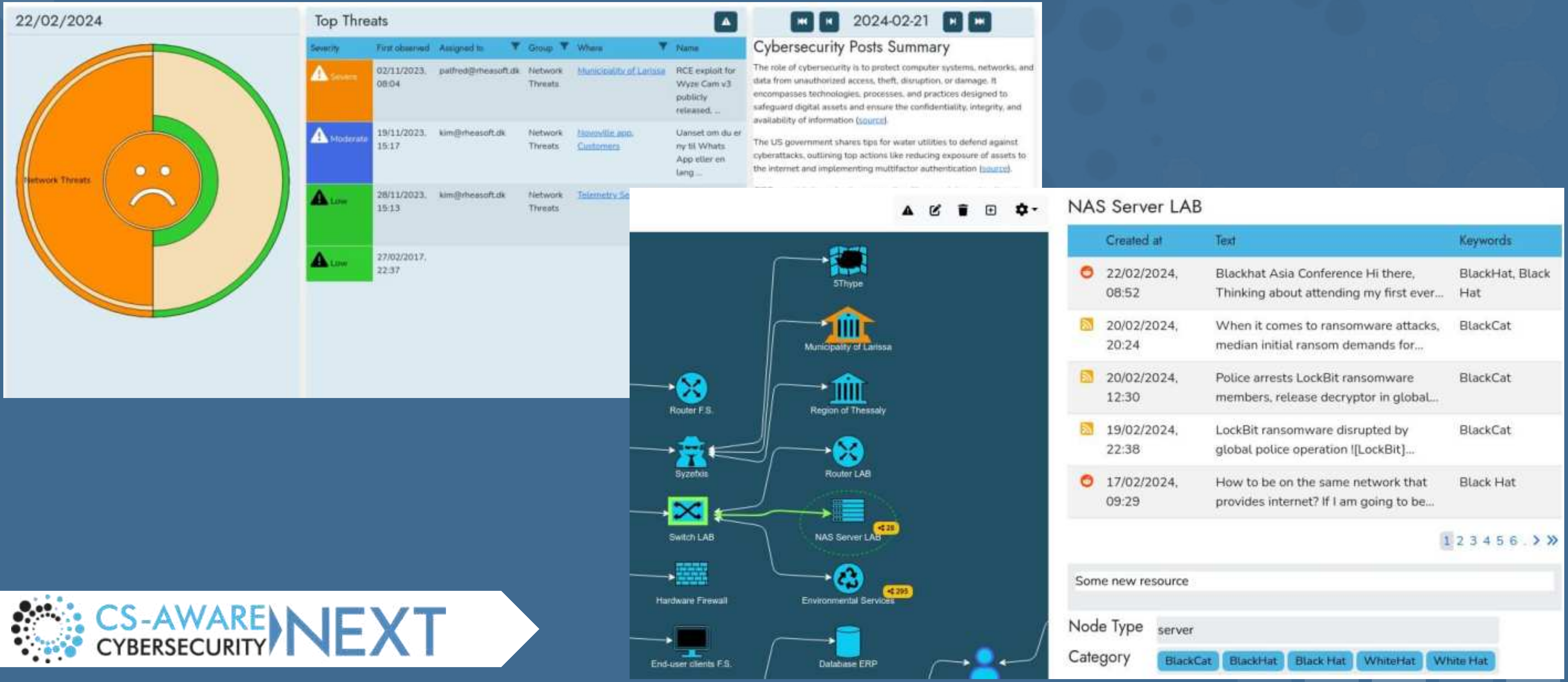
- 2 pillars
  - **Analysis**: Socio-technical workshops allow the people working in an organisation to specify their systems, and how they work in practice (co-design)
  - **Monitoring:** Provide the technical capability for data-driven real-time monitoring and managing the system as specified by the people of an organisation

CS-AWARE process – Analysis





CS-AWARE process – Monitoring



Top Threats

Severity	First observed	Assigned to	Group	Where	Name
Severe	02/11/2023, 08:04	palfred@rheasoft.dk	Network Threats	Municipality of Larissa	RCE exploit for Wyze Cam v3 publicly released. ...
Moderate	19/11/2023, 15:17	kim@rheasoft.dk	Network Threats	Novoville app, Customers	Uanset om du er ny til Whats App eller en lang ...
Low	28/11/2023, 15:13	kim@rheasoft.dk	Network Threats	Telemetry Se	
Low	27/02/2017, 22:37				

2024-02-21

Cybersecurity Posts Summary

The role of cybersecurity is to protect computer systems, networks, and data from unauthorized access, theft, disruption, or damage. It encompasses technologies, processes, and practices designed to safeguard digital assets and ensure the confidentiality, integrity, and availability of information (source).

The US government shares tips for water utilities to defend against cyberattacks, outlining top actions like reducing exposure of assets to the internet and implementing multifactor authentication (source).

NAS Server LAB

Created at	Text	Keywords
22/02/2024, 08:52	Blackhat Asia Conference Hi there, Thinking about attending my first ever...	BlackHat, Black Hat
20/02/2024, 20:24	When it comes to ransomware attacks, median initial ransom demands for...	BlackCat
20/02/2024, 12:30	Police arrests LockBit ransomware members, release decryptor in global...	BlackCat
19/02/2024, 22:38	LockBit ransomware disrupted by global police operation ![LockBit]...	BlackCat
17/02/2024, 09:29	How to be on the same network that provides internet? If I am going to be...	Black Hat

1 2 3 4 5 6 . › »

Some new resource

Node Type: server

Category: BlackCat, BlackHat, Black Hat, WhiteHat, White Hat

CS-AWARE CYBERSECURITY NEXT

CS-AWARE commercialization

- A start-up was founded by core partners of the CS-AWARE project
  - Common exploitation of IP developed during project
  - Common exploitation allows establishing strong branding
  - Development to market readiness
  - Establishing sales channels

- Challenges
  - A TRL7 pilot demonstrator is not a market ready product, additional funding is required for commercialisation
  - Attracting funding at such an early commercial state is challenging
  - We are trying to fundamentally change how organisations manage cybersecurity. Not easy to convince customers, even if offering may be superior.

- We are now in a market ready state with the CS-AWARE lite platform, and are hopeful to have our first customers in 2024

Why CS-AWARE-NEXT?

- Core insight from CS-AWARE project
  - Improving cybersecurity within an organisation is not enough, as organisations are part of ecosystems

- Inter-organisational collaboration
  - Using the CS-AWARE platform as the basis for organisational cybersecurity management, how can we support collaboration among organisations (e.g. supply chain), with a focus on improving regional collaboration
  - Collaboration does not simply happen because it is required, it needs support and focus to develop norms and practices

CS-AWARE-NEXT Objectives

- The project has 8 objectives, clearly linked to the project work plan
  - Data-driven inter-organisational collaboration as the overarching theme
  - Enabling dynamic and pro-active cybersecurity management

CS-AWARE-NEXT Objectives

Objective 1

Dynamic policy support

WP1

Ecosystem Collaboration

AI based data correlation

BC/DR and self-healing

Information sharing

Objective 6

Benchmarking and Profiling

WP7

Objective 7

Implementation, Integration and deployment

WP6

Objective 8

Design science based project implementation and validation

WP7

Design-science based project implementation

> In the design-science paradigm, knowledge and understanding of a problem domain and its solution are achieved in the building and application of the designed artifact
>
> -- Hevner et. al., Design Science in Information Systems Research

Design–science based project implementation

- Design and implement information system artifacts

- **Co-design:** build around user needs, together with the users

- **Iterative design and validation:** Validate various states from conceptualisation to implementation with end users, and adopt according to insights gained

Tools for supporting regional cybersecurity collaboration

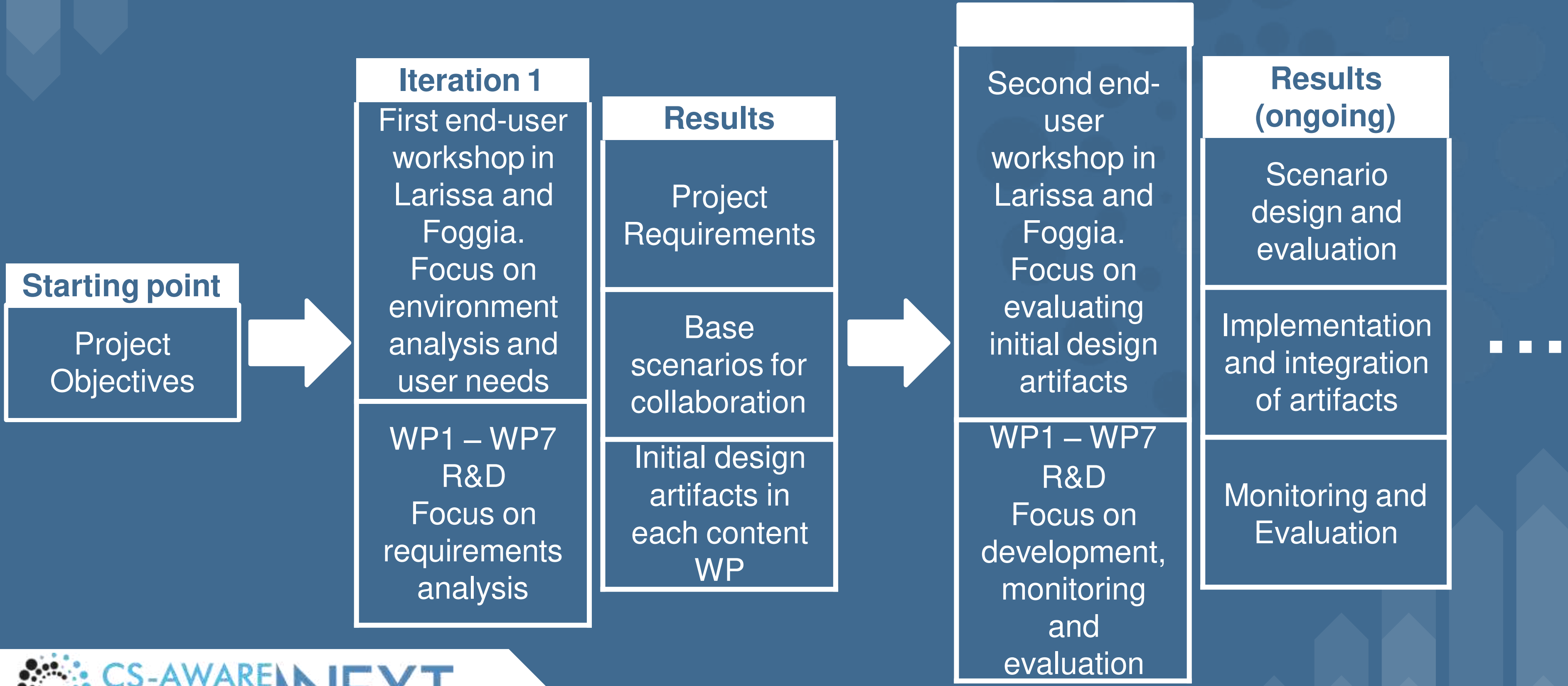
CS-AWARE PLATFORM for information and awareness

- Organisational level
  - Policy support
  - New AI-based data and information engine
  - BC/DR support, self-healing
  - Information sharing with authorities

CS-CONNECT as collaborative interface

- Ecosystem level collaboration
  - Mattermost-based
  - Co-designed collaboration scenarios
- Data-driven
  - Integration with data from
  - CS-AWARE platform Availability of AI-based contextualized data

Project implementation overview



Starting point

Project Objectives

Iteration 1

First end-user workshop in Larissa and Foggia. Focus on environment analysis and user needs

WP1 – WP7 R&D Focus on requirements analysis

Results

Project Requirements

Base scenarios for collaboration

Initial design artifacts in each content WP

Second end-user workshop in Larissa and Foggia. Focus on evaluating initial design artifacts

WP1 – WP7 R&D Focus on development, monitoring and evaluation

Results (ongoing)

Scenario design and evaluation

Implementation and integration of artifacts

Monitoring and Evaluation

...

Two representative pilot regions

Pilot regions reflect current state in Europe well

- Larissa region
  - NIS sectors focus
  - Medium to high cybersecurity maturity
  - Pre-existing collaborations (e.g. health sector)
  - **Driver:** European legislation
  - High understanding of benefits of collaboration, and high motivation work together

- Foggia region
  - Industry focus
  - Low technology use and thus low cybersecurity maturity
  - **Driver:** customer demands
  - Reservations about inter-organisational collaboration (e.g. competitor rivalry)

Smart Risk Management for Business Continuity in Critical Infrastructure

DYNABIC

tecnalia beawre

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070455.

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

The Challenge

Unforeseen **cascading impacts** within critical infrastructures, and inadequate management of business interruptions, leads to **financial losses in the millions and reputational damage**

Ineffective risk management due to lack of business awareness results in misguided risk mitigation efforts and diminishes decision-makers' control.

Lack of standard-based technical security testing makes training of Business disruption risk managers long and expensive.

RISKM4BC

RISKM4BC is a dynamic business risk management framework. Designed to offer both design and operational support, this tool is specifically tailored for cascading impact assessment and real-time risk quantification within the chain of Critical Infrastructures (CIs).

Cascading Effects	Control risks related to **cascading effects** among multiple interrelated CIs.
Prediction on workflows	**Prediction of unwanted incidents and risk** to meet deadlines in your **workflows** to mitigate business disruptions.
Continuous Risk Control	**Bowtie Smart Center** to automate risk identification and predict threats.
NL-enabled	Automated **link between business goals and lower-level system risks** leveraging LLMs.

Technology

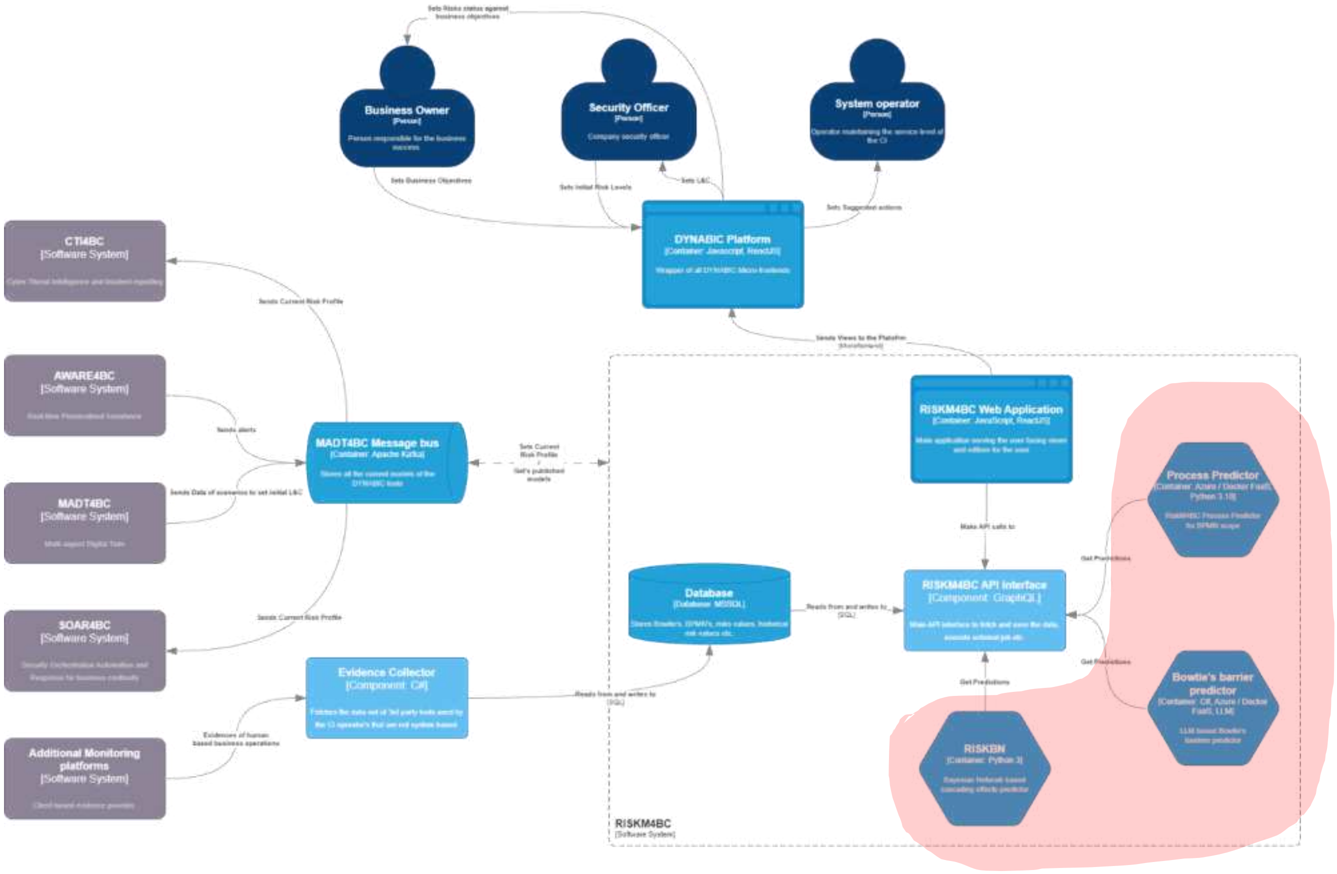
Key technical features:

- Automated **risk likelihood and impact learning**.
- Risk propagation through **cascading effects** calculation.
- Workflow digital twins for **workflow execution prediction**.
- Live preventive and mitigative action **recommendation in natural Language**.
- Automated link of **business goals** with system risks.

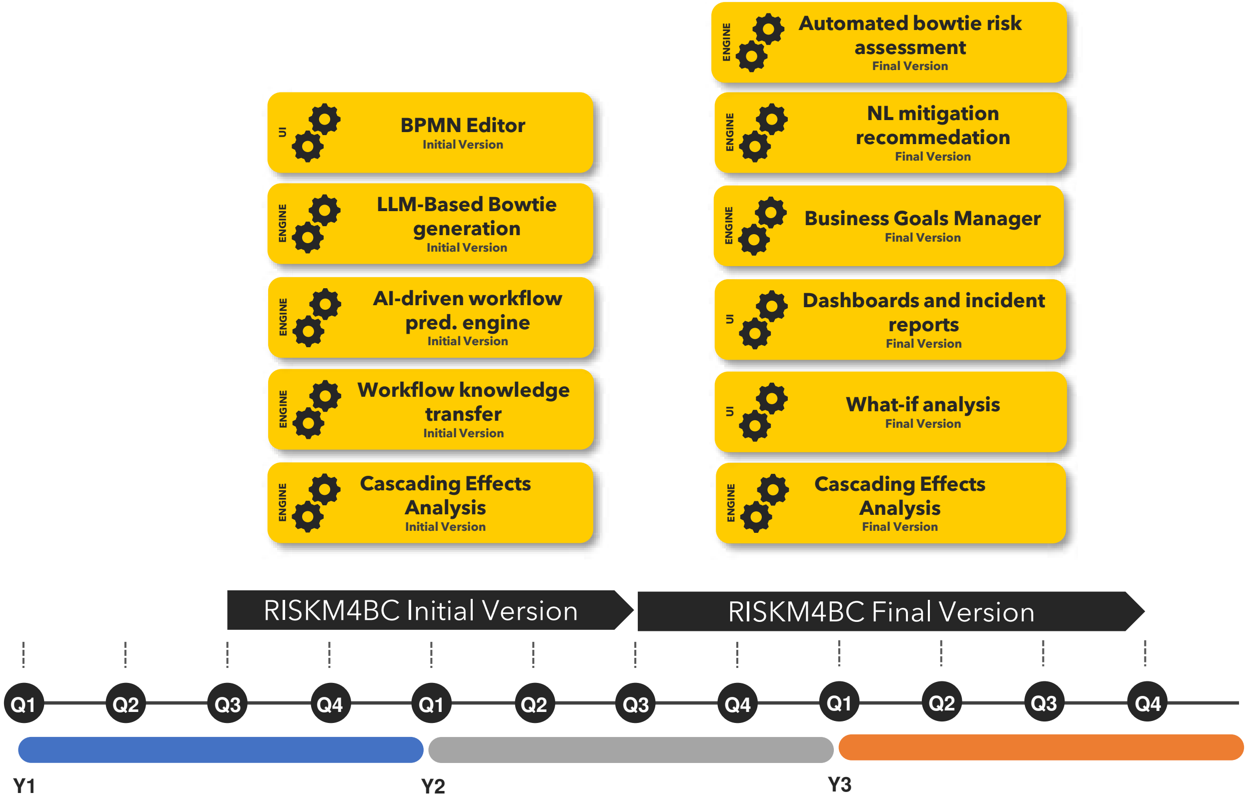
Proprietary predictive AI algorithms (transformers, LSTM, …)

Advanced use of the latest innovations in generative AI (LLM)

RISKM4BC Context within DYNABIC


Sets Risks status against business objectives
Business Owner
[Person]
Person responsible for the business success
Security Officer
[Person]
Company security officer
System operator
[Person]
Operator maintaining the service level of the CI
Sets Business Objectives
Sets Initial Risk Levels
Sets L&C
Sets Suggested actions
DYNABIC Platform
[Container: Javascript, ReactJS]
Wrapper of all DYNABIC Micro frontends
CTI4BC
[Software System]
Cyber Threat Intelligence and Incident reporting
Sends Current Risk Profile
Sends Views to the Platform
[Microfrontend]
AWARE4BC
[Software System]
Real-time Personalized Assistance
Sends alerts
MADT4BC Message bus
[Container: Apache Kafka]
Stores all the current models of the DYNABIC tools
Sets Current Risk Profile
/
Get's published models
MADT4BC
[Software System]
Multi aspect Digital Twin
Sends Data of scenarios to set initial L&C
RISKM4BC Web Application
[Container: JavaScript, ReactJS]
Main application serving the user facing views and editions for the user
Process Predictor
[Container: Azure / Docker FaaS, Python 3.10]
RiskM4BC Process Predictor for BPMN scope
Make API calls to
Get Predictions
SOAR4BC
[Software System]
Security Orchestration Automation and Response for business continuity
Sends Current Risk Profile
Database
[Database: MSSQL]
Stores Bowtie's, BPMN's, risks values, historical risk values etc.
Reads from and writes to
[SQL]
RISKM4BC API interface
[Component: GraphQL]
Main API interface to fetch and save the data, execute schedule job etc.
Get Predictions
Evidence Collector
[Component: C#]
Fetches the data out of 3rd party tools used by the CI operator's that are not system based
Reads from and writes to
[SQL]
Bowtie's barrier predictor
[Container: C#, Azure / Docker FaaS, LLM]
LLM based Bowtie's barriers predictor
Get Predictions
Evidences of human based business operations
Additional Monitoring platforms
[Software System]
Client based evidence provider
RISKBN
[Container: Python 3]
Bayesian Network based cascading effects predictor
RISKM4BC
[Software System]

Current status and technical roadmap



UI — **BPMN Editor** — Initial Version

ENGINE — **LLM-Based Bowtie generation** — Initial Version

ENGINE — **AI-driven workflow pred. engine** — Initial Version

ENGINE — **Workflow knowledge transfer** — Initial Version

ENGINE — **Cascading Effects Analysis** — Initial Version

ENGINE — **Automated bowtie risk assessment** — Final Version

ENGINE — **NL mitigation recommedation** — Final Version

ENGINE — **Business Goals Manager** — Final Version

UI — **Dashboards and incident reports** — Final Version

UI — **What-if analysis** — Final Version

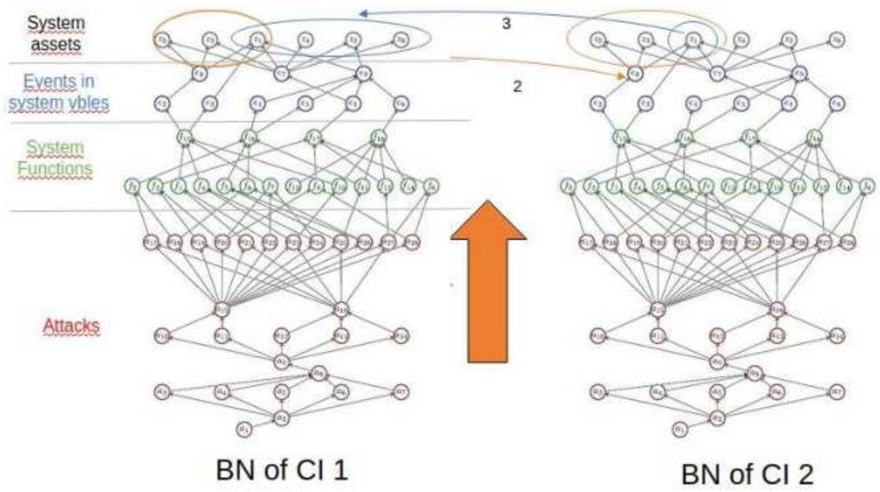
ENGINE — **Cascading Effects Analysis** — Final Version

RISKM4BC Initial Version

RISKM4BC Final Version

RISKM4BC Research goals

Research Objective 1: Create a system to control risks related to cascading effects among multiple interrelated CIs.

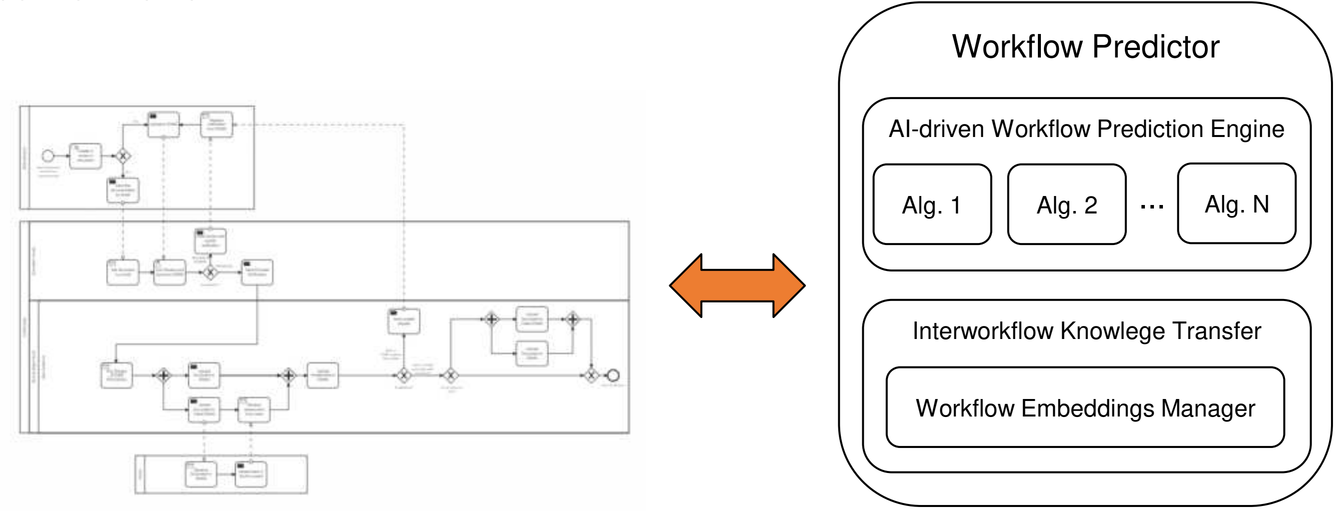


- Targeting CI interdependencies and Cascading Effects with other CIs.
- High degree of granularity on each CI, effects on different parts on connected CIs.
- CIs do not share their risk models (and live with it)

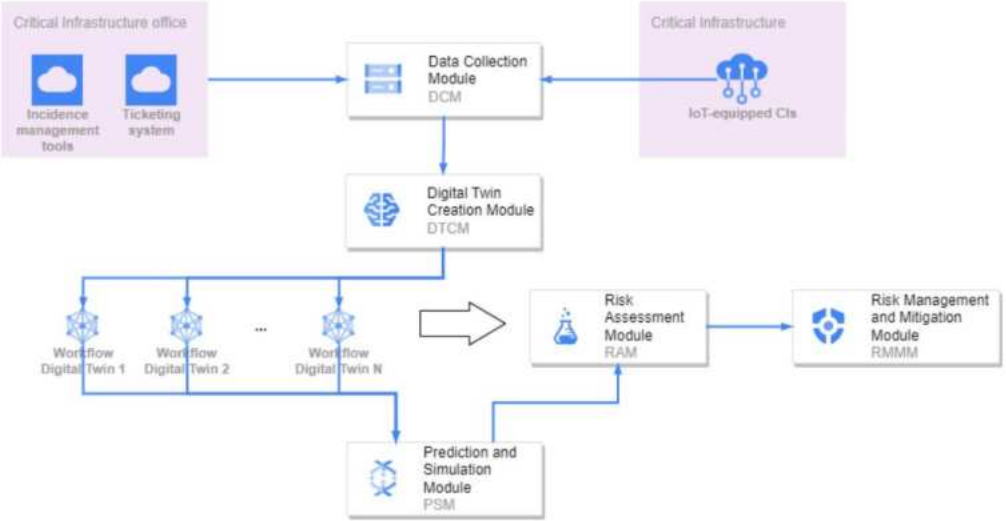
RISKM4BC Research goals

Research Objective 2: Create a live risk matrix leveraging AI-driven workflow digital twins to predict workflow evolution probability distributions

- Create and compare different AI-driven predictive algorithms to forecast current workflows evolution
- Create an ensemble approach with weighted algorithms
- Enable prediction capacity in non-observed workflows through knowledge transfer from different past observed workflows


Workflow Predictor
AI-driven Workflow Prediction Engine
Alg. 1
Alg. 2
…
Alg. N
Interworkflow Knowlege Transfer
Workflow Embeddings Manager

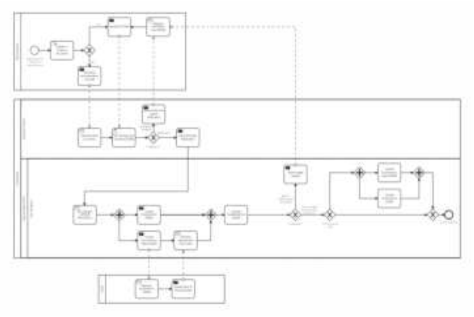
Continuous Risk Management Concept

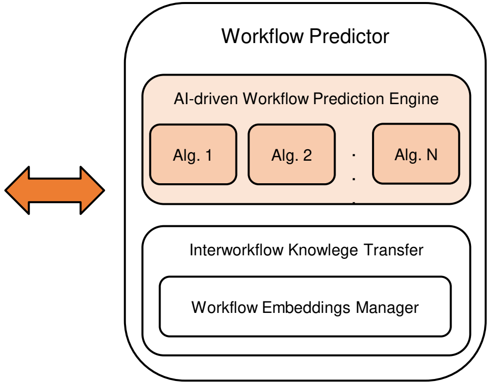


WP3 Research goals

Research Objective 2: Create a live risk matrix leveraging AI-driven workflow digital twins to predict workflow evolution probability distributions

- Create and compare different AI-driven predictive algorithms to forecast current workflows evolution
- Create an ensemble approach with weighted algorithms
- Enable prediction capacity in non-observed workflows through knowledge transfer from different past observed workflows




Workflow Predictor

AI-driven Workflow Prediction Engine

Alg. 1

Alg. 2

Alg. N

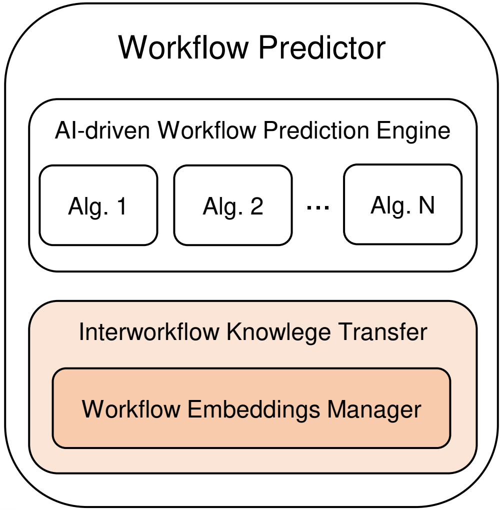
Interworkflow Knowlege Transfer

Workflow Embeddings Manager

WP3 Research goals

Research Objective 2: Create a live risk matrix leveraging AI-driven workflow digital twins to predict workflow evolution probability distributions

- Create and compare different AI-driven predictive algorithms to forecast current workflows evolution
- Create an ensemble approach with weighted algorithms
- Enable prediction capacity in non-observed workflows through knowledge transfer from different past observed workflows


Workflow Predictor

AI-driven Workflow Prediction Engine

Alg. 1

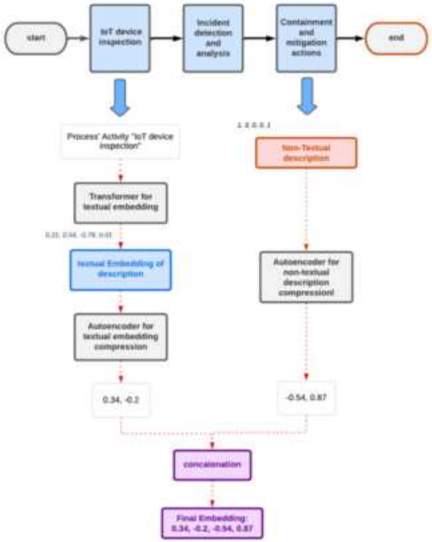
Alg. 2

…

Alg. N

Interworkflow Knowlege Transfer

Workflow Embeddings Manager


start
IoT device inspection
Incident detection and analysis
Containment and mitigation actions
end
Process' Activity "IoT device inspection"
Transformer for textual embedding
0.21, 0.94, -0.78, 0.61
textual Embedding of description
Autoencoder for textual embedding compression
0.34, -0.2
1, 2, 0, 0, 1
Non-Textual description
Autoencoder for non-textual description compression!
-0.54, 0.87
concatenation
Final Embedding: 0.34, -0.2, -0.54, 0.87

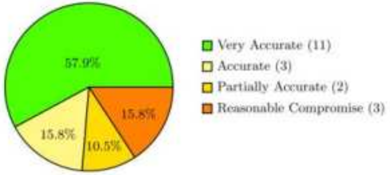
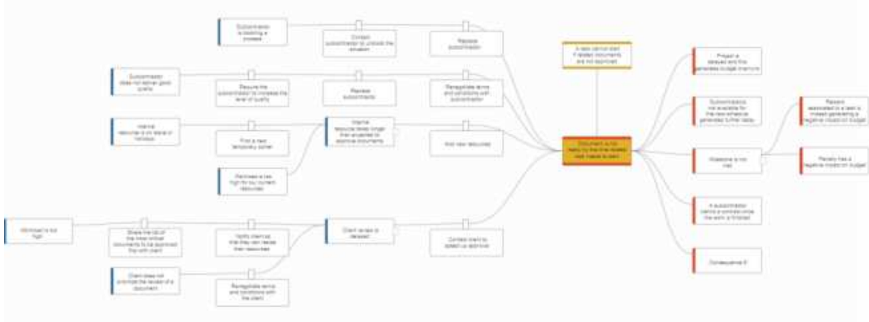


Fig. 9. Summary of Process Embedding Evaluations

WP3 Research goals

Research Objective 3: Create a bowtie smart center able to automate risk identification, predict threats and risk evolution, and chained risk effects

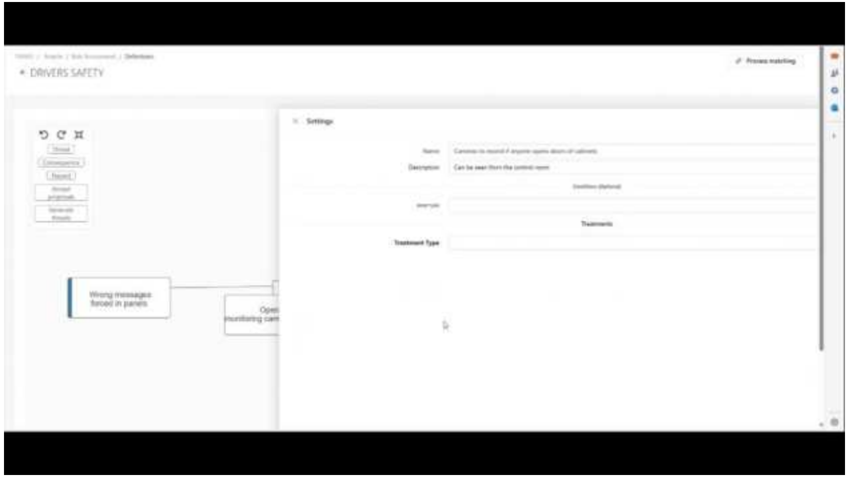
- Create mechanisms to automatically create bowties out of existing workflows, initial textual descriptions or partially defined bowties
- Create a continuous self-learning model to predict bowtie activations
- Enable prediction capacity in non-observed bowties through knowledge transfer from different past observed bowties
- Automated link of business objectives with lower-level bowtie components through LLMs



WP3 Research goals

Research Objective 3: Create a bowtie smart center able to automate risk identification, predict threats and risk evolution, and chained risk effects

- Create mechanisms to automatically create bowties out of existing workflows, initial textual descriptions or partially defined bowties
- Create a continuous self-learning model to predict bowtie activations
- Enable prediction capacity in non-observed bowties through knowledge transfer from different past observed bowties
- Automated link of business objectives with lower-level bowtie components through LLMs



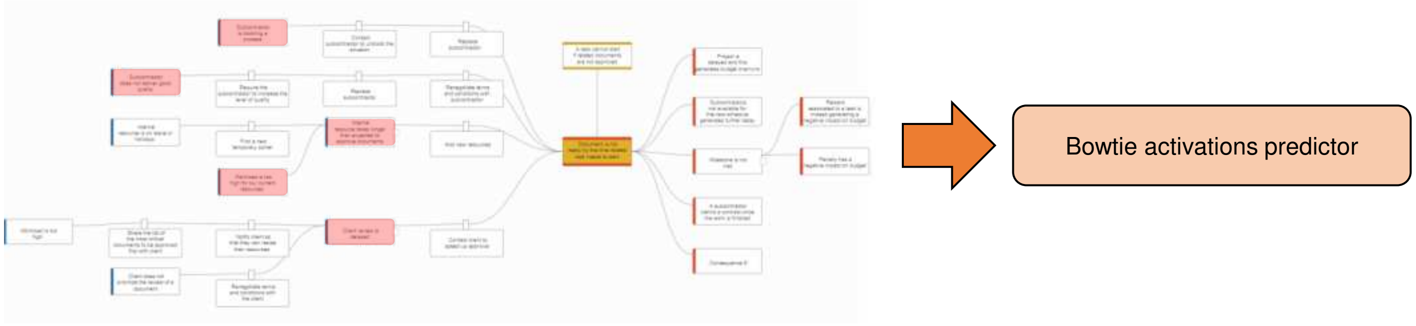
Name: Cameras to record if anyone opens doors of cabinets

Description: Can be seen from the control room

WP3 Research goals

Research Objective 3: Create a bowtie smart center able to automate risk identification, predict threats and risk evolution, and chained risk effects

- Create mechanisms to automatically create bowties out of existing workflows, initial textual descriptions or partially defined bowties
- Create a continuous self-learning model to predict bowtie activations
- Enable prediction capacity in non-observed bowties through knowledge transfer from different past observed bowties
- Automated link of business objectives with lower-level bowtie components through LLMs



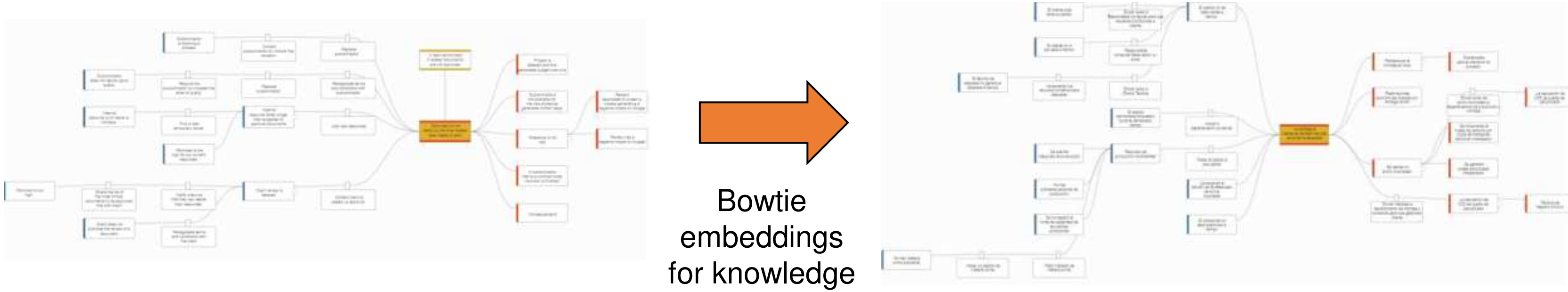
Bowtie activations predictor

t1, t3, t4, t7, t1, t2

WP3 Research goals

Research Objective 3: Create a bowtie smart center able to automate risk identification, predict threats and risk evolution, and chained risk effects

- Create mechanisms to automatically create bowties out of existing workflows, initial textual descriptions or partially defined bowties
- Create a continuous self-learning model to predict bowtie activations
- Enable prediction capacity in non-observed bowties through knowledge transfer from different past observed bowties
- Automated link of business objectives with lower-level bowtie components through LLMs

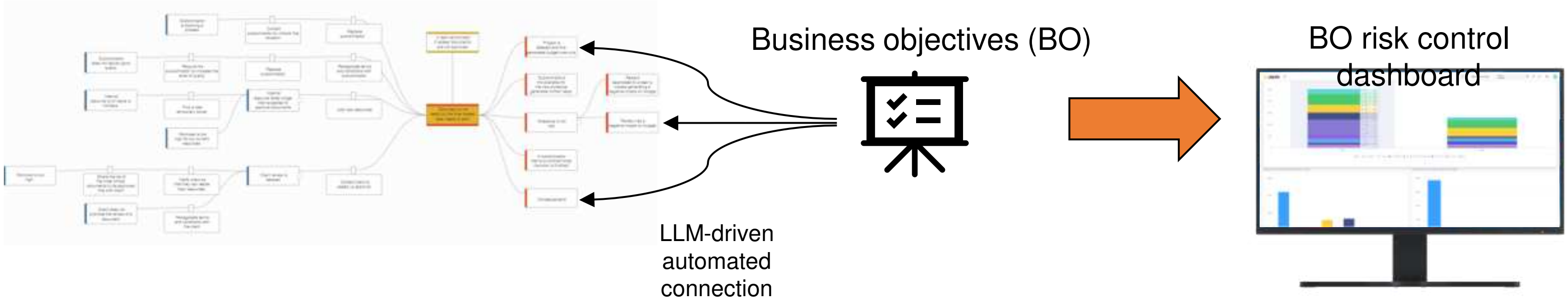


Bowtie embeddings for knowledge transfer

WP3 Research goals

Research Objective 3: Create a bowtie smart center able to automate risk identification, predict threats and risk evolution, and chained risk effects

- Create mechanisms to automatically create bowties out of existing workflows, initial textual descriptions or partially defined bowties
- Create a continuous self-learning model to predict bowtie activations
- Enable prediction capacity in non-observed bowties through knowledge transfer from different past observed bowties
- Automated link of business objectives with lower-level bowtie components through LLMs



Business objectives (BO)

BO risk control dashboard

LLM-driven automated connection

Q&A

ANSWERS TO YOUR QUESTIONS

CS-AWARE CYBERSECURITY NEXT AI4CYBER DYNABIC

AI Supporting Cyber Risks And Resilience Of Critical Infrastructures.

Insights From CS-AWARE-NEXT, DYNABIC, & AI4CYBER

Online Webinar

30 September 2024

16:00 - 17:30

Funded by the European Union

AI4CYBER

TRUSTWORTHY ARTIFICIAL INTELLIGENCE FOR CYBERSECURITY REINFORCEMENT AND SYSTEM RESILIENCE

Trustworthy AI for cybersecurity solutions

Webinar “AI supporting Cyber Risks and Resilience of Critical Infrastructures”

30/09/2024

Erkuden Rios, AI4CYBER Project Manager
erkuden.rios@tecnalia.com

tecnalia
MEMBER OF BASQUE RESEARCH
& TECHNOLOGY ALLIANCE

Funded by the European Union

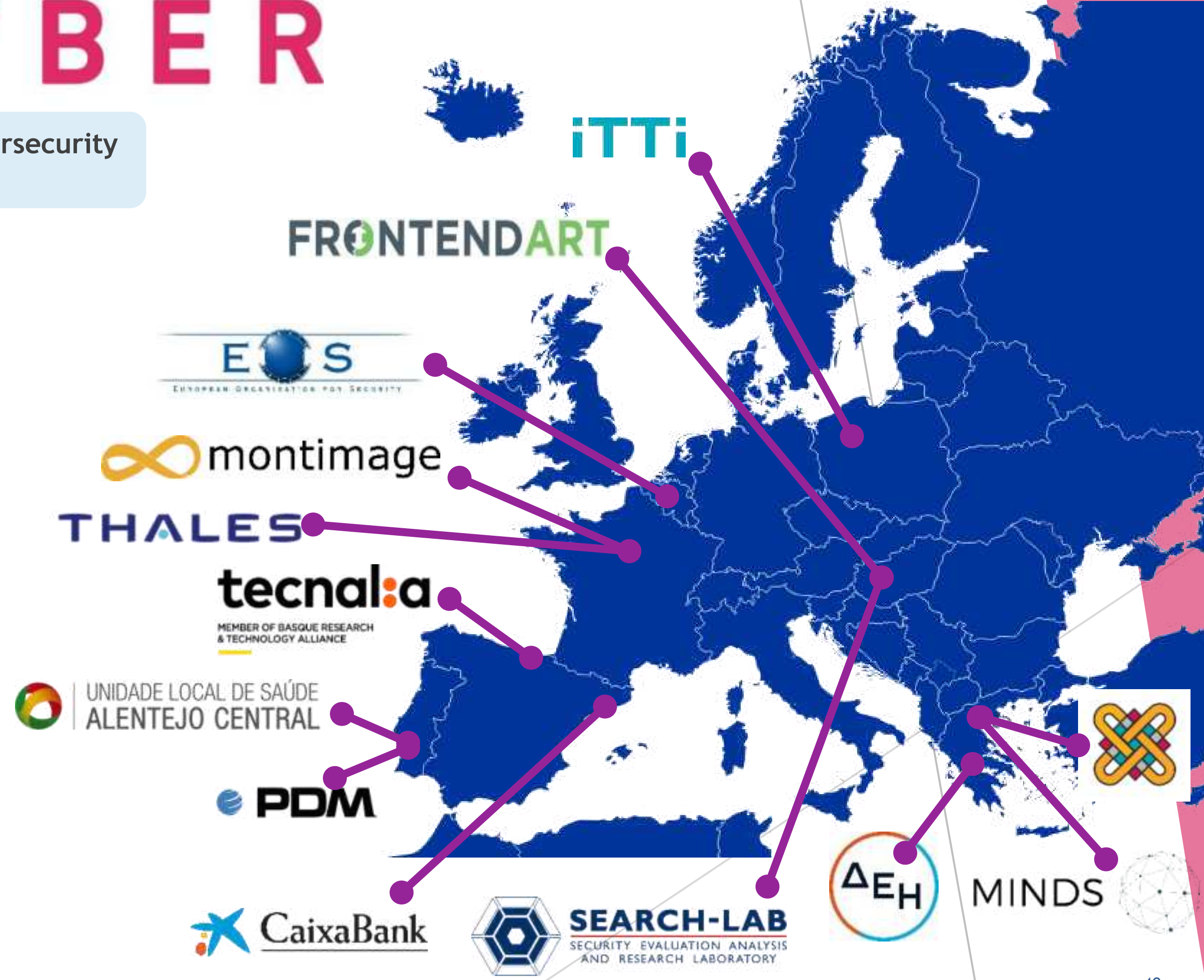
This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101070450.

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

AI4CYBER

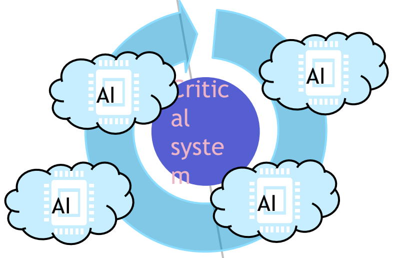
Trustworthy Artificial Intelligence for Cybersecurity Reinforcement and System Resilience

- Grant Agreement ID: 101070450
- Project Type: RIA
- Project Coordinator: Tecnalia
- Consortium: 13 partners
- Budget: € 3.998.413,00 €
- Start Date: 01/09/2022
- Duration: 3 years

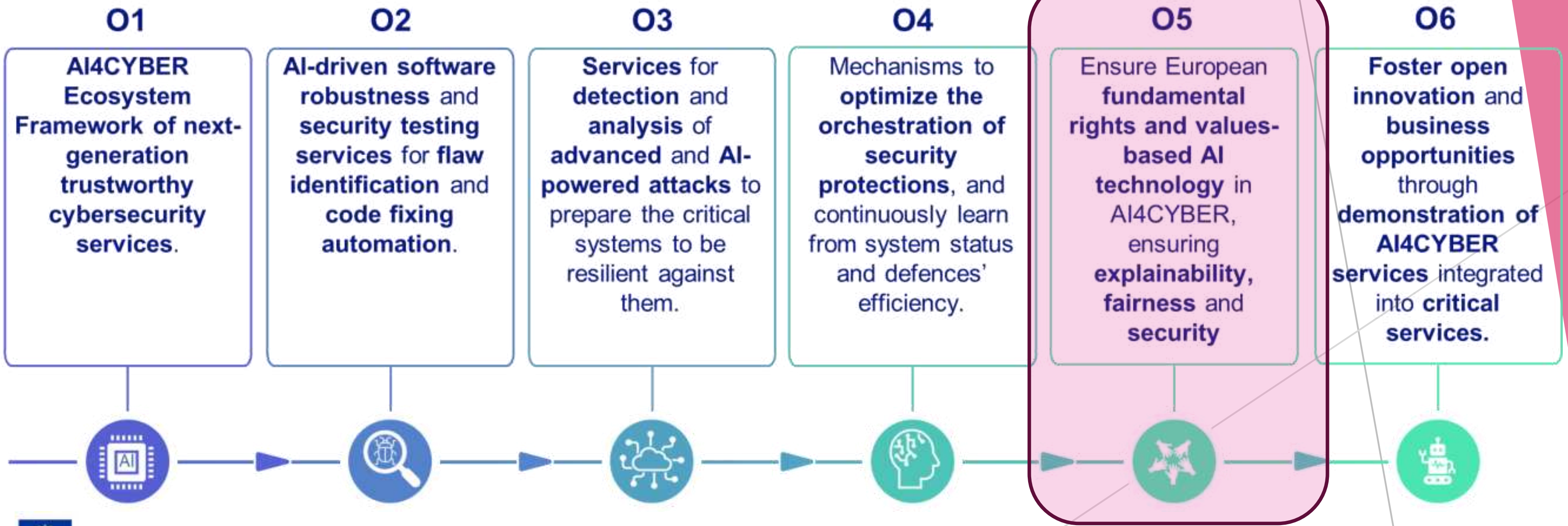

iTTi
FRONTENDART
EOS
European Organisation for Security
montimage
THALES
tecnalia
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
UNIDADE LOCAL DE SAÚDE ALENTEJO CENTRAL
PDM
CaixaBank
SEARCH-LAB
SECURITY EVALUATION ANALYSIS AND RESEARCH LABORATORY
ΔEH
MINDS

Funded by the European Union

Key objectives


Continuum of care

To establish an **Ecosystem Framework of next generation AI-based services** for supporting critical system developers and operators to **efficiently manage** system **robustness**, **resilience**, and appropriate **response** in the face of **advanced and AI-powered cyberattacks**.



O1

AI4CYBER Ecosystem Framework of next-generation trustworthy cybersecurity services.

O2

AI-driven software robustness and **security testing services** for **flaw identification** and **code fixing automation**.

O3

Services for **detection** and **analysis** of **advanced** and **AI-powered attacks** to prepare the critical systems to be resilient against them.

O4

Mechanisms to **optimize the orchestration of security protections**, and continuously learn from system status and defences' efficiency.

O5

Ensure European **fundamental rights and values-based AI technology** in AI4CYBER, ensuring **explainability, fairness** and **security**

O6

Foster open innovation and **business opportunities** through **demonstration of AI4CYBER services** integrated into **critical services.**

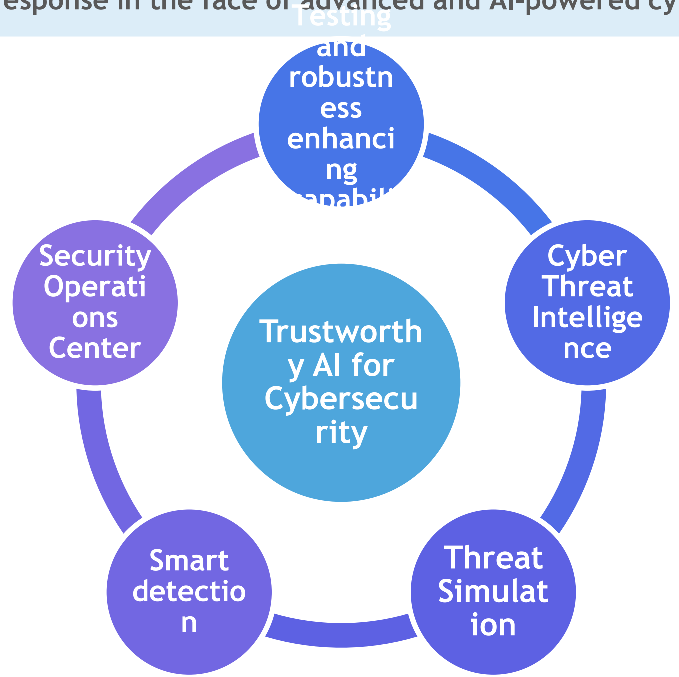
Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

AI4CYBER in a nutshell

Establishing an **Ecosystem Framework** of next generation **AI-based services** for critical system **robustness, resilience**, and appropriate **response in the face of advanced and AI-powered cyberattacks**.

11 Key Results that cover **6 cybersecurity areas**



Testing and robustness enhancing capabil

Cyber Threat Intelligence

Threat Simulation

Smart detection

Security Operations Center

Trustworthy AI for Cybersecurity

3 Demonstrators that foster **innovation**

Energy

Banking

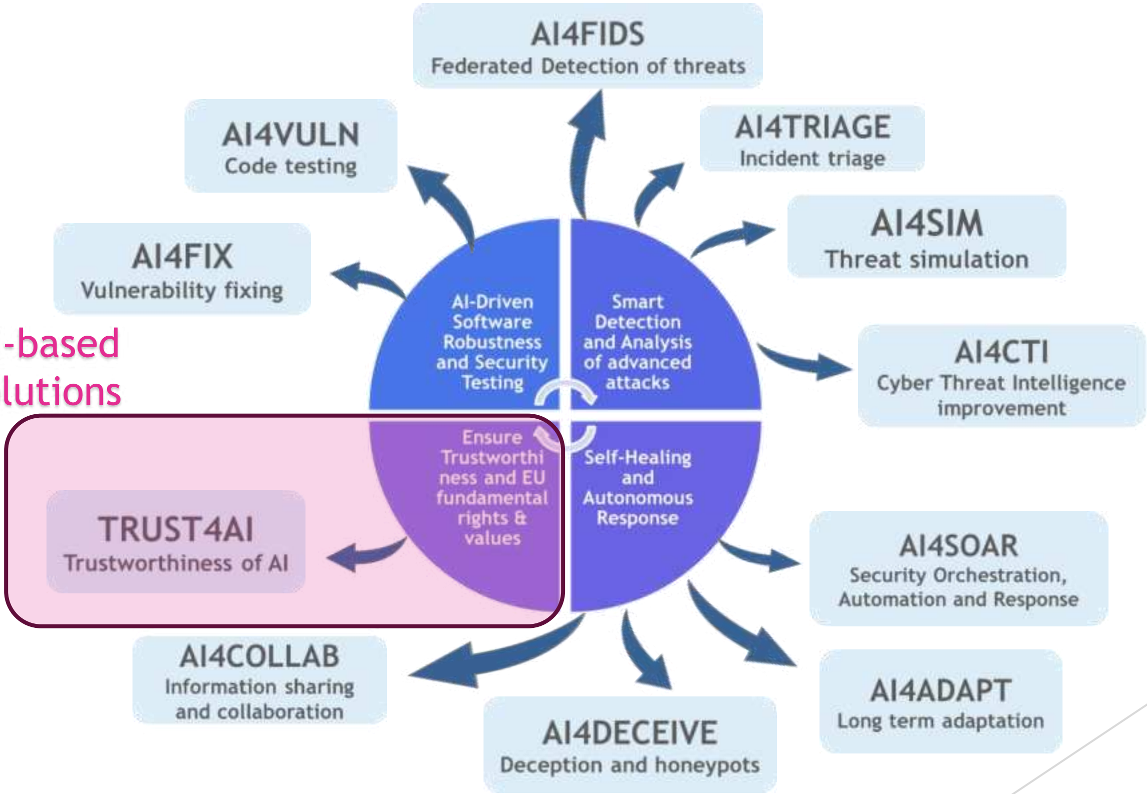
Health

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

AI4CYBER Framework

TRUSWORTHY AI-based cybersecurity solutions



AI4FIDS
Federated Detection of threats

AI4VULN
Code testing

AI4TRIAGE
Incident triage

AI4SIM
Threat simulation

AI4FIX
Vulnerability fixing

AI4CTI
Cyber Threat Intelligence improvement

AI-Driven Software Robustness and Security Testing

Smart Detection and Analysis of advanced attacks

Ensure Trustworthiness and EU fundamental rights & values

Self-Healing and Autonomous Response

TRUST4AI
Trustworthiness of AI

AI4SOAR
Security Orchestration, Automation and Response

AI4COLLAB
Information sharing and collaboration

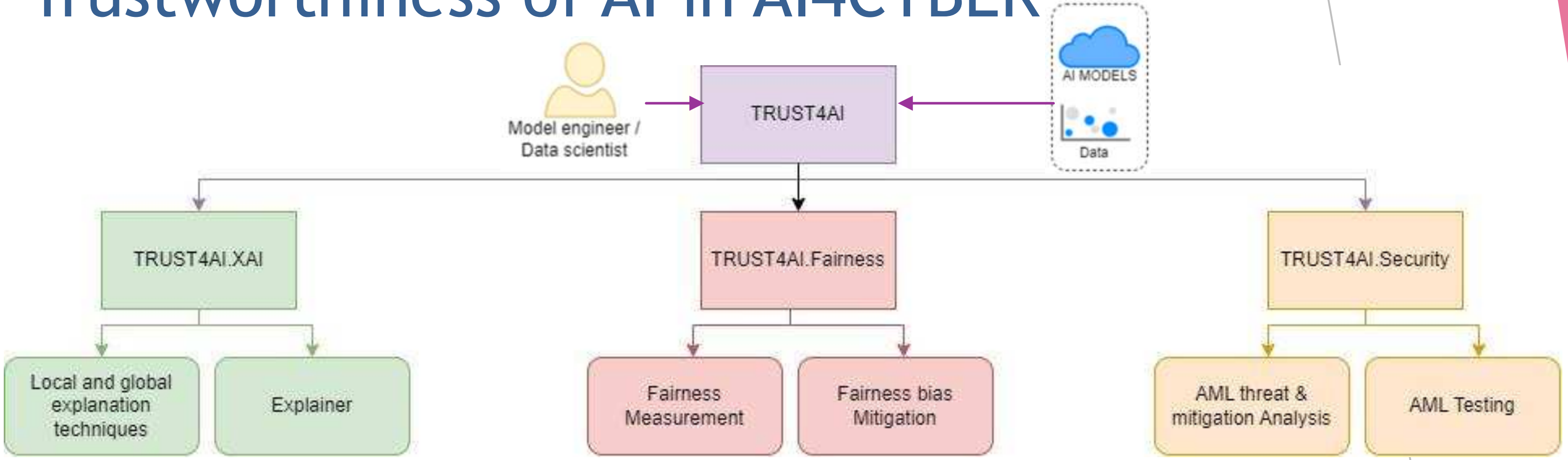
AI4DECEIVE
Deception and honeypots

AI4ADAPT
Long term adaptation

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

Trustworthiness of AI in AI4CYBER



Explainability – XAI (i.e. Interpretability) of ML/AI models, to allow better understanding of the model to the data scientist who develops the ML model or to a potential end-user.

Fairness of the ML/AI models, to allow to correct potential bias against some sensitive attributes or against sub-populations.

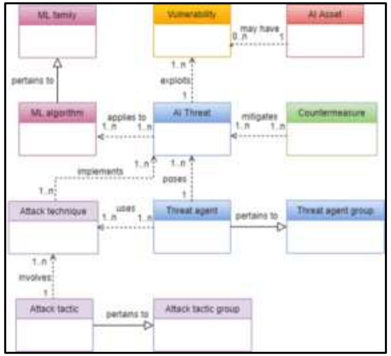
Security of the ML/AI models, to allow to learn potential Adversarial Machine Learning attacks and to protect against them.

Funded by the European Union

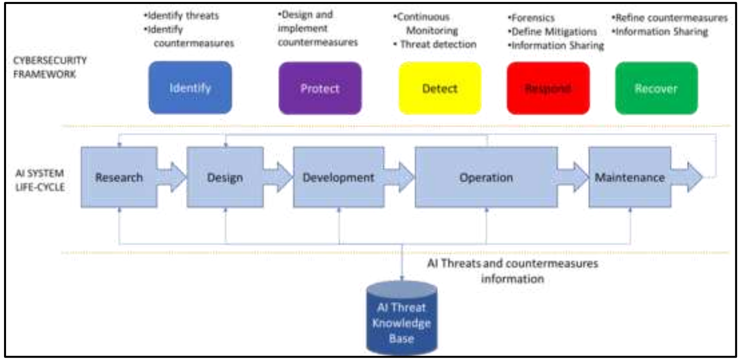
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

AI Security landscape

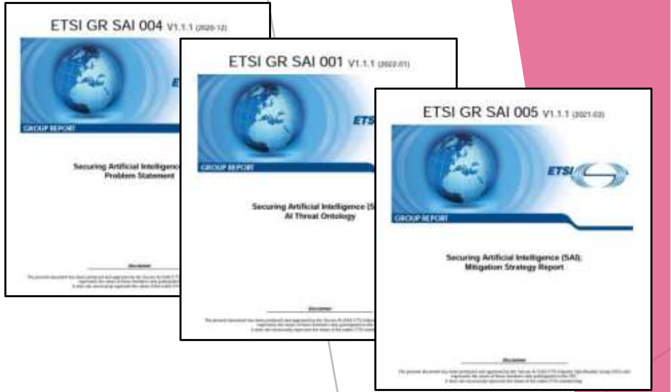
AI4CYBER AI Threat model



SPARTA AI Threat KB



ETSI SAI


ETSI GR SAI 004 V1.1.1 (2020-12)
GROUP REPORT
Securing Artificial Intelligence
Problem Statement
ETSI GR SAI 001 V1.1.1 (2022-01)
GROUP REPORT
Securing Artificial Intelligence (S
AI Threat Ontology
ETSI GR SAI 005 V1.1.1 (2021-03)
ETSI
GROUP REPORT
Securing Artificial Intelligence (SAI);
Mitigation Strategy Report

ENISA

SECURING MACHINE LEARNING ALGORITHMS
AI
MULTILAYER FRAMEWORK FOR GOOD CYBERSECURITY PRACTICES FOR AI
ARTIFICIAL INTELLIGENCE AND CYBERSECURITY RESEARCH
CYBERSECURITY AND PRIVACY IN AI – FORECASTING DEMAND ON ELECTRICITY GRIDS

MITRE ATLAS

ATLAS™

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below.

* indicates an adaptation from ATT&CK. Click on links to learn more about each item, or view ATLAS tactics and techniques using the links at the top navigation bar.

NIST AI 100-2e2023 ipd
AML taxonomy

NIST AI 100-2e2023 ipd
Adversarial Machine Learning
A Taxonomy and Terminology of Attacks and Mitigations
NIST

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

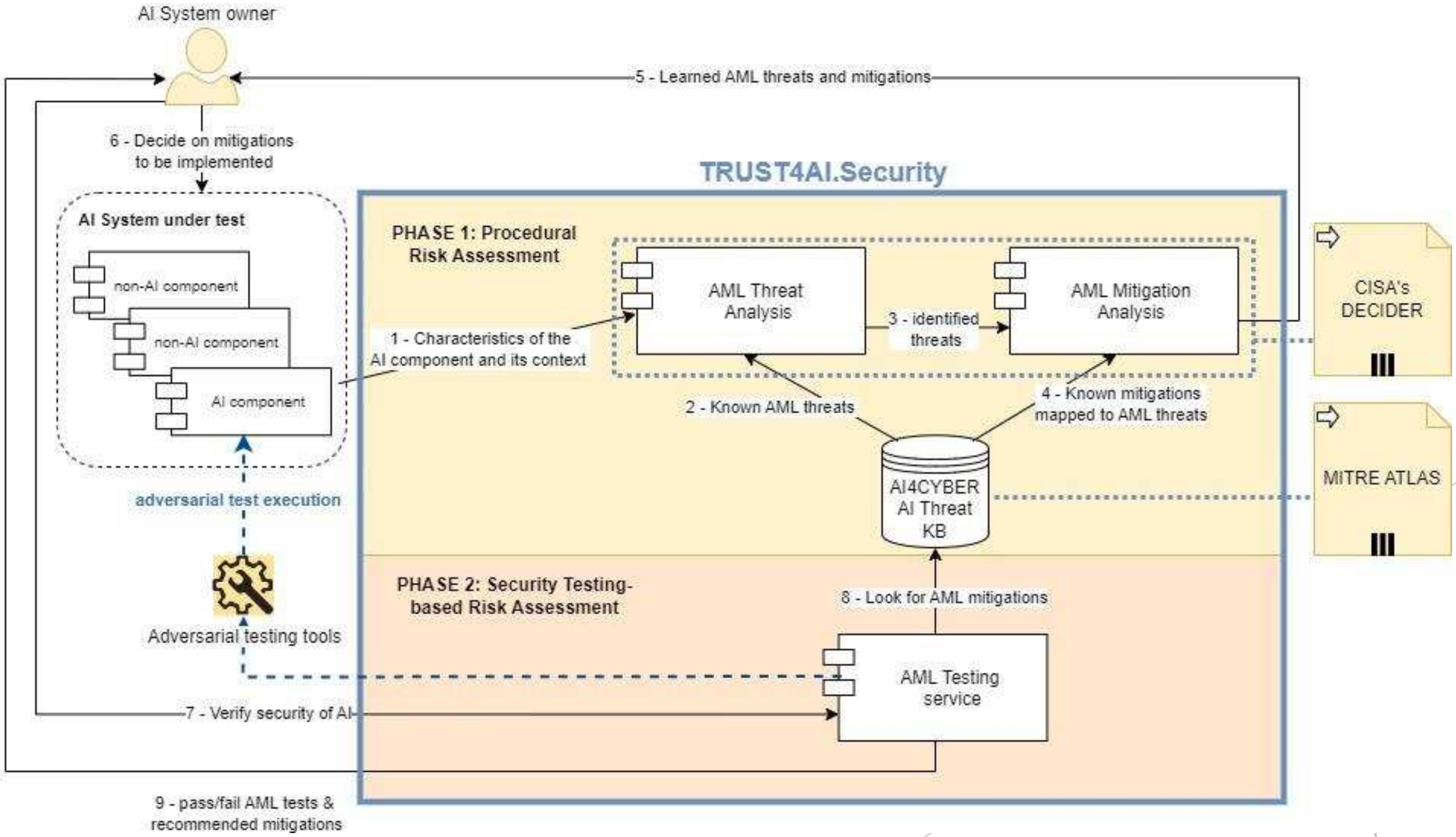
TRUST4AI.Security Objectives

- Security of AI as part of Robustness of AI.
- Protect models from Adversarial Machine Learning (AML) attacks, including: ***data protection***, ***poisoning***, ***evasion***, and ***oracle (privacy) attacks***.
- Offer mechanisms for **countering** attacks to the AI-based cybersecurity tools, e.g., intrusion and anomaly detection tools based on AI models.
- Builds on top of the *AI Threat and Countermeasure Knowledge Base* from H2020 SPARTA project and MITRE ATLAS.
- Leverages CISA's Decider tool.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

AI system risk assessment with TRUST4AI.Security



AI System owner

5 - Learned AML threats and mitigations

6 - Decide on mitigations to be implemented

TRUST4AI.Security

AI System under test

non-AI component

non-AI component

AI component

PHASE 1: Procedural Risk Assessment

1 - Characteristics of the AI component and its context

AML Threat Analysis

3 - identified threats

AML Mitigation Analysis

2 - Known AML threats

4 - Known mitigations mapped to AML threats

AI4CYBER AI Threat KB

CISA's DECIDER

MITRE ATLAS

adversarial test execution

Adversarial testing tools

PHASE 2: Security Testing-based Risk Assessment

8 - Look for AML mitigations

AML Testing service

7 - Verify security of AI

9 - pass/fail AML tests & recommended mitigations

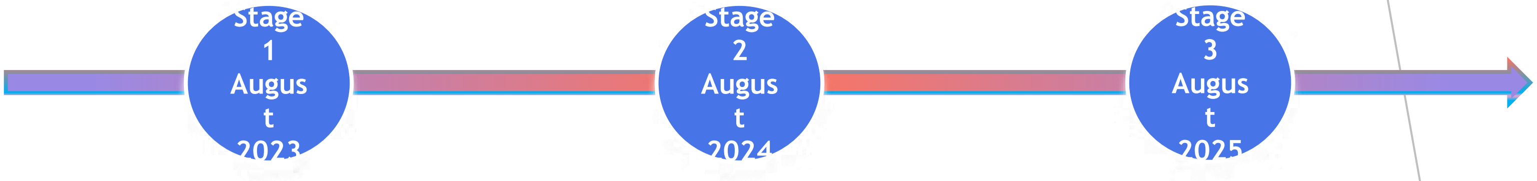
Funded by the European Union

This prc
under g

AI4CYBER

AI4CYBER Timeline

Project implementation
September 2022 – August 2025



Stage 1
August 2023

Stage 2
August 2024

Stage 3
August 2025

Stage 1 - Design
Initial architecture of the AI4CYBER Framework

Stage 2 - First release
First release of the AI4CYBER Framework services and tools

Stage 3 - Final release
Final release of the AI4CYBER Framework services and tools

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

TRUST4AI publications

- Kurek, W., Pawlicki, M., Pawlicka, A., Kozik, R., & Choraś, M. (2023, July). **Explainable Artificial Intelligence 101: Techniques, Applications and Challenges.** In International Conference on Intelligent Computing (pp. 310-318). Singapore: Springer Nature Singapore (SPRINGER LNCS).
- Pawlicki, M.: **Towards Quality Measures for xAI algorithms: Explanation Stability,** DSAA2023 (CORE A)
- Pawlicki, M., Pawlicka, A., Śrutek, M., Kozik, R., Choraś, M. **Interpreting Intrusions - The Role of Explainability in AI-based Intrusion Detection Systems** (IP&C 2023)
- Thouvenot V., Huynh C.B. **TSCFKit and CFKit, two Python modules dedicated to counterfactual generation**, JDS 2023.
- Dr. Marek Pawlicki panelist in CLAIRE AQuA: Cybersecurity of AI and AI for Cybersecurity. (https://www.youtube.com/watch?v=u44CiZhkbnY)
- Prof. Michał Choraś keynote on "Trustworthy and Explainable AI (xAI) in Emerging Network Security Applications" in ARES2023 (https://www.ares-conference.eu/workshops-eu-symposium/ens-2023/)
- Uccello, F., Pawlicki, M., D'Antonio, S., Kozik, R., & Choraś, M. (2024, April). **A Novel Approach to the Use of Explainability to Mine Network Intrusion Detection Rules.** In Asian Conference on Intelligent Information and Database Systems (pp. 70-81). Singapore: Springer Nature Singapore.
- Pawlicki, M., Puchalski, D., Szelest, S., Pawlicka, A., Kozik, R., & Choraś, M. (2024, July). **Introducing a Multi-Perspective xAI Tool for Better Model Explainability.** In Proc. of the 19th International Conference on Availability, Reliability and Security (pp. 1-8).
- Kozik, R., Kątek, G., Gackowska, M., Kula, S., Komorniczak, J., Ksieniewicz, P., ... & Choraś, M. (2024). **Towards explainable fake news detection and automated content credibility assessment: Polish internet and digital media use-case.** *Neurocomputing*, *608*, 128450.

TRUST4AI publications

- Pawlicki, M. (2024). **ARIA, HaRIA and GeRIA: Novel Metrics for Pre-Model Interpretability**. IEEE Access.
- Nguyen, M. D., Bouaziz, A., Valdes, V., Rosa Cavalli, A., Mallouli, W., & Montes De Oca, E. (2023, August). A **deep learning anomaly detection framework with explainability and robustness.** In Proc. of the 18th International Conference on Availability, Reliability and Security (pp. 1-7).
- Asimopoulos, D. C., Radoglou-Grammatikis, P., Makris, I., Mladenov, V., Psannis, K. E., Goudos, S., & Sarigiannidis, P. (2023, August)**. Breaching the defense: Investigating fgsm and ctgan adversarial attacks on IEC 60870-5-104 AI-enabled intrusion detection systems**. In Proc. of the 18th International Conference on Availability, Reliability and Security (pp. 1-8).
- Villarini, B., Radoglou-Grammatikis, P., Lagkas, T., Sarigiannidis, P., & Argyriou, V. (2023, July). **Detection of Physical Adversarial Attacks on Traffic Signs for Autonomous Vehicles**. In 2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT) (pp. 31-37).
- Bouaziz, A., Nguyen, M. D., Valdés, V., Cavalli, A. R., & Mallouli, W. (2023, July). **Study on Adversarial Attacks Techniques, Learning Methods and Countermeasures: Application to Anomaly Detection**. In ICSOFT (pp. 510-517).
- Iturbe, E., Rios, E., & Toledo, N. (2023, December). **Towards trustworthy Artificial Intelligence: Security risk assessment methodology for Artificial Intelligence systems.** In 2023 IEEE International Conference on Cloud Computing Technology and Science (CloudCom) (pp. 291-297).
- Blog post “Trustworthy AI” - General overview of TRUST4AI https://ai4cyber.eu/?p=1136

AI4CYBER

TRUSTWORTHY ARTIFICIAL INTELLIGENCE FOR CYBERSECURITY REINFORCEMENT AND SYSTEM RESILIENCE

https://ai4cyber.eu/

https://x.com/Ai4Cyber

https://www.linkedin.com/company/ai4cyber/

Erkuden Rios

AI4CYBER Project Coordinator

erkuden.rios@tecnalia.com

Thank you for your attention!

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

CS-AWARE CYBERSECURITY NEXT · AI4CYBER · DYNABIC

AI Supporting Cyber Risk Management And Resilience Of Critical Infrastructures;

Insights From CS-AWARE-NEXT, DYNABIC, & AI4CYBER

Online Webinar

30 September 2024

16:00 - 17:30

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101070450.