Safe and Trusted Human Centric Artificial Intelligence in Future Manufacturing Lines

STAR

www.star-ai.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°956573

USING EXPLAINABLE AI FOR TRUSTED PRODUCTION SYSTEMS

John Soldatos INTRASOFT International

“Explainable Artificial Intelligence in Manufacturing”

Workshop organized by the Cluster of AI in Manufacturing (AI-MAN) Projects

11.10.2021

www.star-ai.eu

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°956573

STAR PROJECT OVERVIEW

- Start date: 1 January 2021
- End date: 31 December 2023
- Overall budget € 5 999 253,75

INTRASOFT INTERNATIONAL

Project Coordinator

THALES

PHILIPS

SIEMENS Ingenuity for life

IBER-OLEFF

GFT

DFKI

Institut "Jožef Stefan" Ljubljana, Slovenija

SUPSI

UNIVERSITY OF PIRAEUS RESEARCH CENTER

UBITECH ubiquitous solutions

R2M RESEARCH TO MARKET SOLUTION

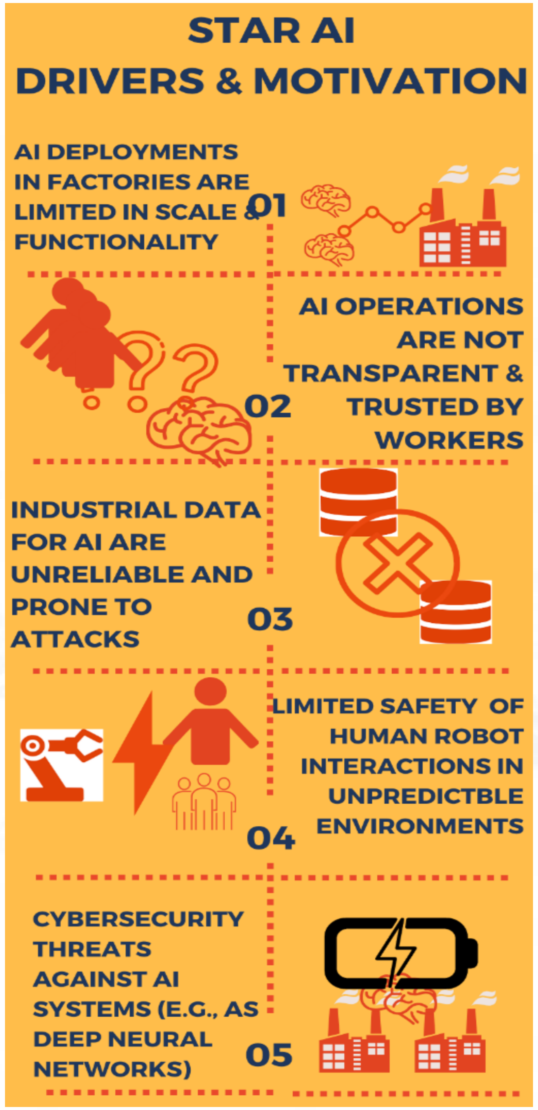
UNPARALLEL

university of groningen

ARTHUR'S LEGAL Challenge the Status Quo

STAR'S MISSION

- Safe, Trusted and Human Centric AI in Manufacturing
- STAR helps manufacturers and industrial automation vendors to build and deploy Safe Reliable and Trusted Human Centric AI systems in real-life manufacturing environments.
- Main Drivers:
  - Enable AI systems to acquire knowledge in order to take timely and safe decisions in dynamic and unpredictable environments.
  - EU HLEG's Ethical Guidelines in Manufacturing Lines (forerunner of AI regulation proposal by European Parliament)



STAR'S WORLD-CLASS CUTTING-EDGE EDGE AI RESEARCH

Explainable AI — Why did you do this?

- Explain to Factory Workers and Quality Engineers the rules and principles of the AI systems operation
- Increasing Transparency and Trust on AI Systems

Active Learning — Robot-to-Human: Is this piece defected?

- Query human where not sure what to do next!
- Accelerate Knowledge Acquisition for AI

Simulated Reality — Shorten Reinforcement Learning Cycle

- Simulate the next actions of Reinforcement Learning than expecting convergence

Human Centric Digital Twins — What-if-Analysis with the Human in Loop

- Simulation & Detection of Safety Issues
- Optimal Deployment of Automated Mobile Robots
- Detection of Safety Zones

(Cyber)Security for AI Systems — Protection of AI Systems against Adversarial Attacks

STAR: ENABLING SAFE, SECURE & ETHICAL AI IN MANUFACTURING

- Explainable & Transparent AI Systems
- Active Learning & Simulated Reality for Human-AI Collaboration
- AI4EU Virtualized Digital Innovation Hub for Safe & Secure AI in Manufacturing
- Cyber Security Solutions for AI Systems in Manufactutirng
- Human-Centric Simulations for Safe AI in Manufacturing

EXPECTED IMPACT

- INCREASED INTELLIGENCE & FLEXIBILITY OF PRODUCTION LINES
- SAFE HUMAN-ROBOT COLLABORATION AT SCALE
- FASTER UPTAKE OF AI SOLUTIONS (QUALITY4.0, CO-BOTS)
- ETHICAL IMPACT IN MANUFACTURING IN-LINE WITH HLEG RECOMMENDATIONS
- RESEARCH (E.G., SIMULATED REALITY, ACTIVE LEARNING, EXPLAINABLE AI) PLACING EU AT FOREFRONT OF GLOBAL AI R&D

XAI: FROM BLACK-BOX AI MODELS TO EXPLAINABLE & INTERPRETABLE MODELS

- Black-box Models (e.g., Deep Learning)
  - Why did you do that?
  - Is there a better option?
  - Is this successful & efficient?
  - Is this a failure?
  - Shall I trust you?
  - When do we get an error?

XAI Models (e.g., LIME, SHAP etc.)

I understand why
I understand why there are no better options
I know when you succeed and when you fail
I know when I can trust you
I know why and when an error occurs

ROLE & USES OF XAI IN STAR

1. Explain AI-based decisions to stakeholders (e.g., workers, plant operators)

2. Use the explanation to perform a task e.g.,
   - Analysis: Identify production process configurations that lead to defects - Using Machine Learning / Deep Learning Explainability
   - Autonomy: Decide which tasks can be undertaken by an autonomous system (e.g., drone or robot) - Using Reinforcement Learning Explainability

3. Generating of Credible Synthetic Data - Data Augmentation

4. Identifying Adversarial Actions and Cybersecurity attacks
   - XAI helps signalling abnormal behaviours

5. Legal & Regulatory Compliance
   - Abide by regulatory principles / mandates e.g., transparency, human oversight etc.
   - HLEG / EU AI Regulatory Compliance

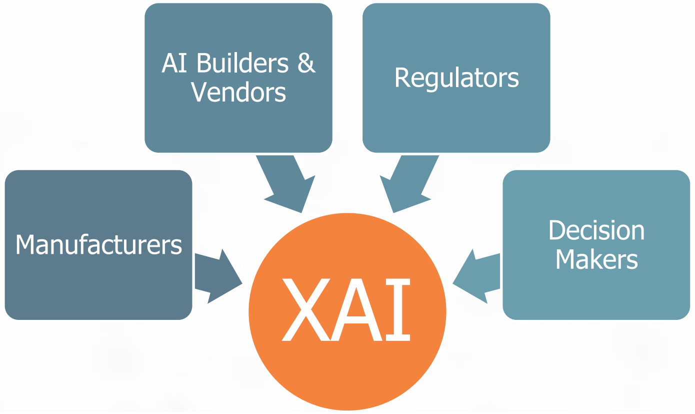
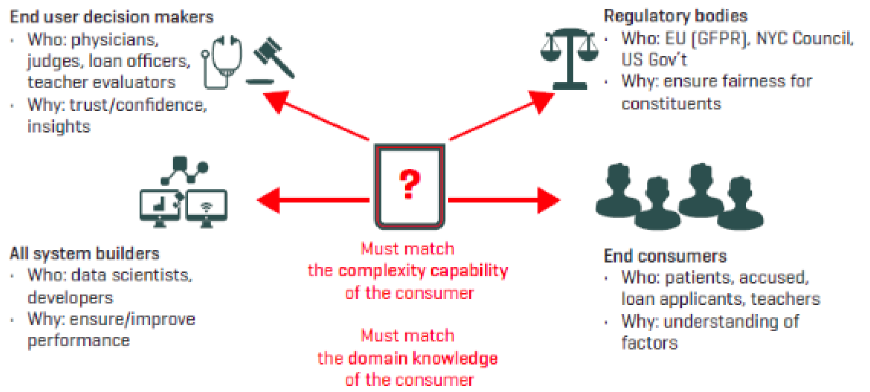
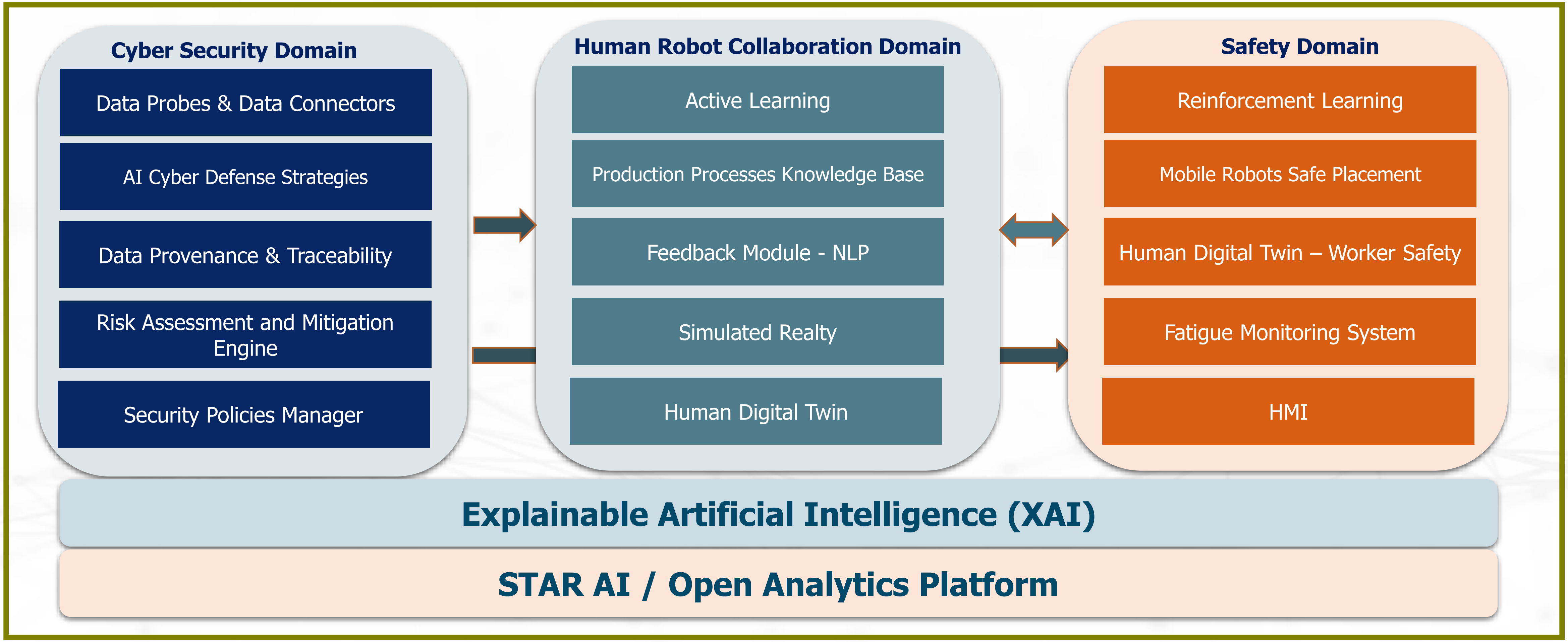


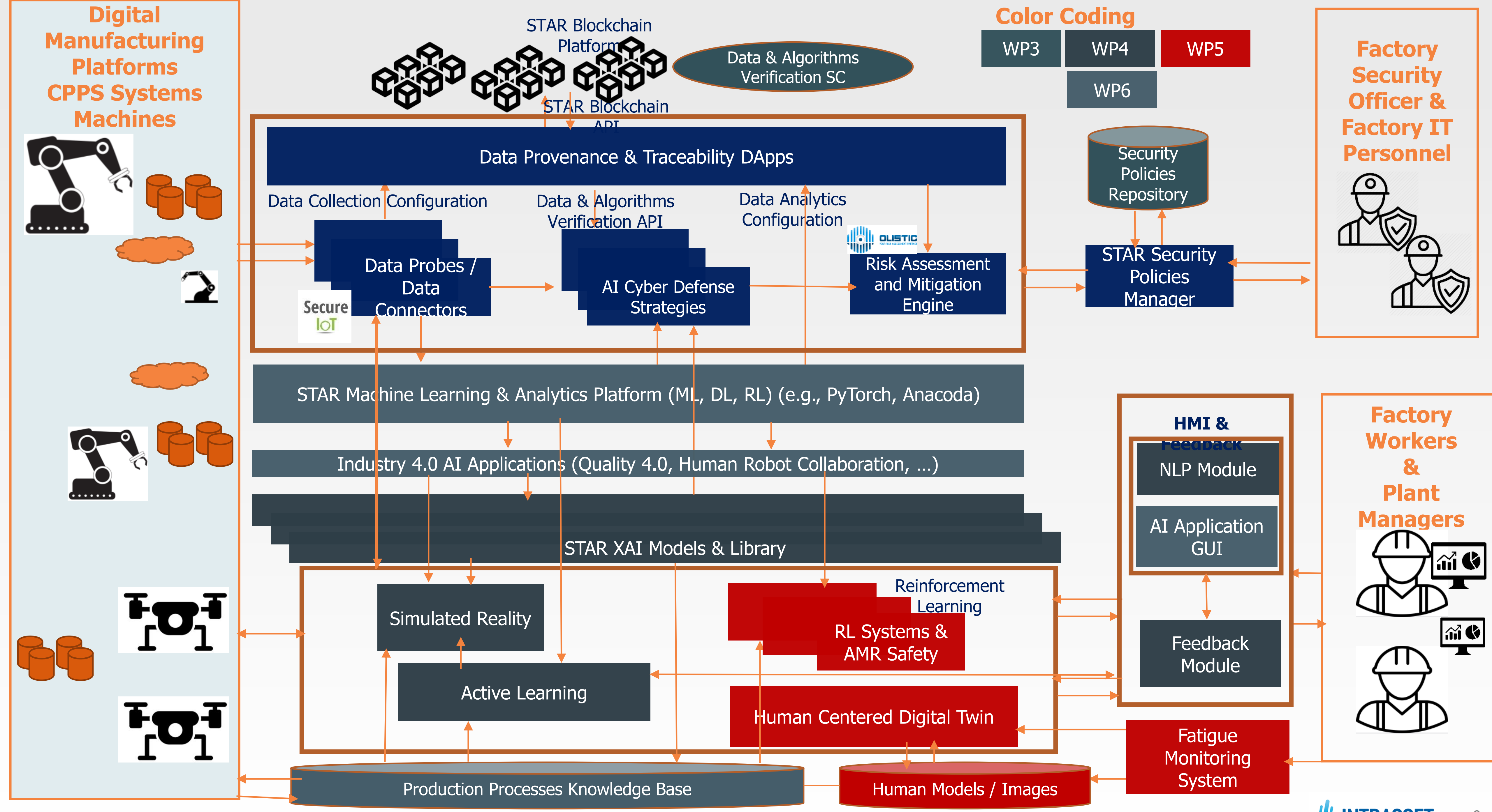
Figure 1. The many groups interested in explainable AI.



Hind, Michael (2019), XRDS: Crossroads, The ACM Magazine for Students — AI and Interpretation, Volume 25 Issue 3, Spring 2019, Pages 16–19

STAR REFERENCE ARCHITECTURE MODEL





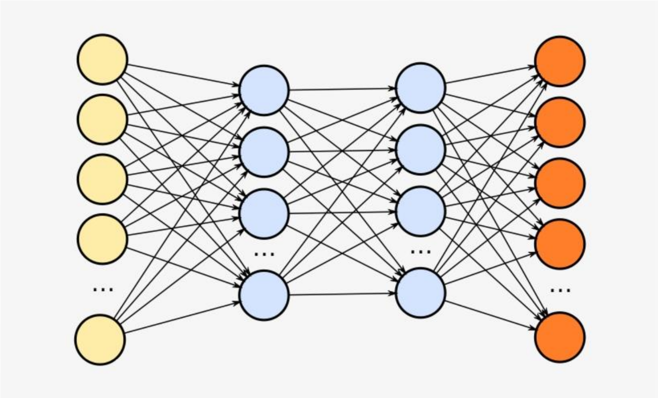
INTRASOFT INTERNATIONAL

THE STAR XAI LIBRARY (1)

- Input Components:
  - AI Algorithms to be explained as black-boxes
  - Specific Instances with predicted classes
  - Access to the internal architecture of the models
- Output Components:
  - Different kinds of explanations
  - Visualized explanations
- Goal: Produce explainable models (e.g., white-glass) without compromising performance

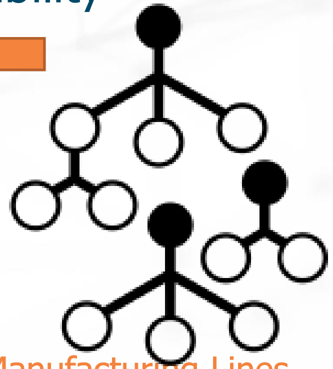
Deep Neural Networks: High Accuracy, Low Explainability

Deep Explainability



Decision Trees – Random Forests: Low-Medium Accuracy, Medium-High Explainability

Interpretable Models



THE STAR XAI LIBRARY (2)

- Implementation of explainability algorithms:
  - Layer Wise Relevance Propagation (LRP) variations
  - Prediction Difference Analysis (PDA) variations
  - LIME variations
  - etc
- Visualize the outcomes of algorithms
- Fit complex Deep Learning models to simpler interpretable ones:
  - Fit classification models to interpretable ones (decision trees etc)
  - Extract models to define human interpretable rules
- Present the above methods to the human factor to boost transparency of the deployed models

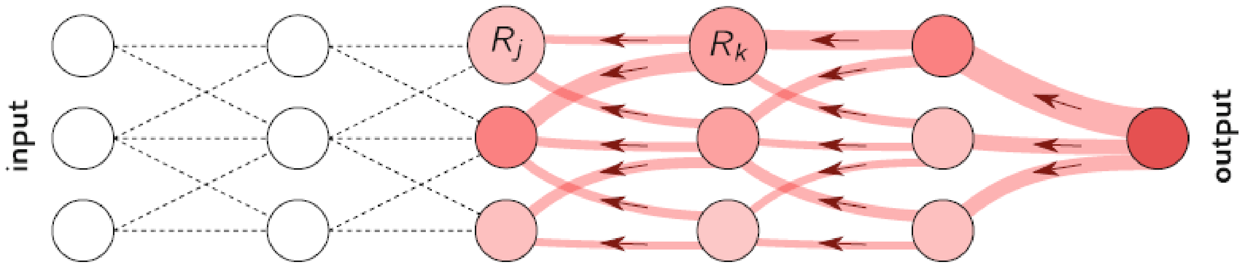
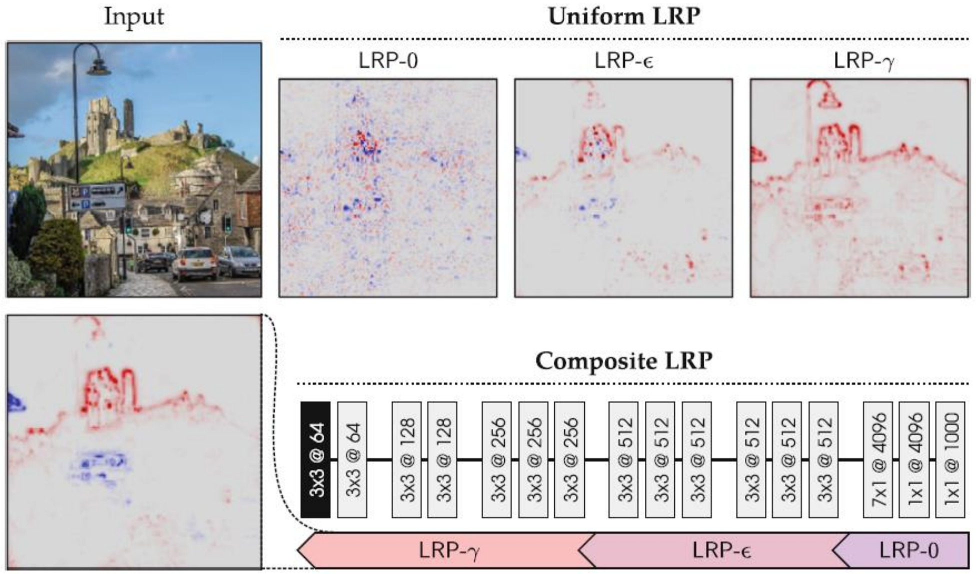


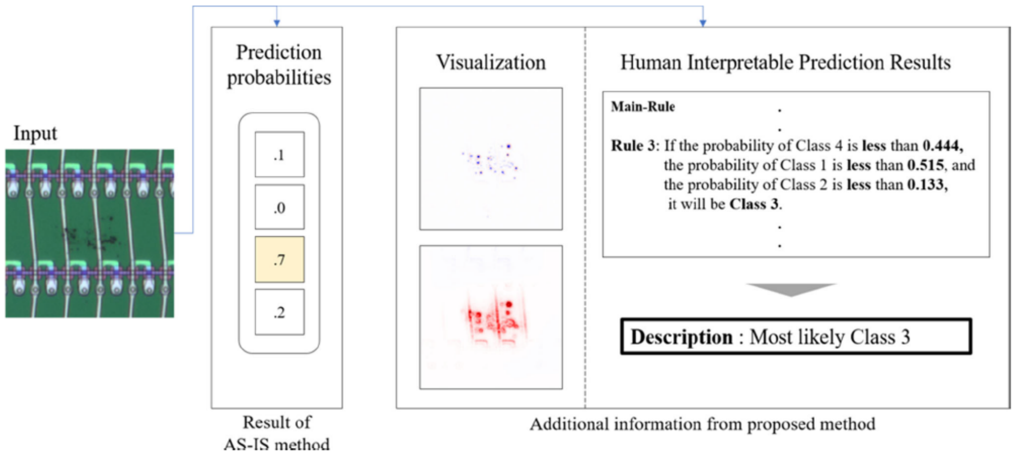
Illustration of the LRP procedure

Montavon G., Binder A., Lapuschkin S., Samek W., Müller KR. (2019) Layer-Wise Relevance Propagation: An Overview. In: Samek W., Montavon G., Vedaldi A., Hansen L., Müller KR. (eds)



UC1. EXPLAINING DECISIONS: XAI FOR QUALITY INSPECTION

- Explanations of classification models
  - Image data + Attribution methods
  - Produce attribution maps + Visualize into heatmaps
  - Highlight features responsible for or against the predicted class
- Model-agnostic methods
  - Applied to different models
  - Produce more general solutions
  - Example: LRP variant (local interpretability) + rules
- Evaluating the Quality of explanations
  - Time complexity -> produce real time results
  - Produce human interpretable explanations



UC2. EXPLAINING DECISIONS: XAI FOR MOBILE ROBOTS OPTIMIZATION

- Explaining human-robot interactions
  - XAI for Deep Reinforcement Learning
    - Transparent algorithms
    - Post-hoc explainability
      - Analysis after the RL agent finishes training and execution.
      - Most post-hoc methods used on visual inputs like images.
      - Saliency methods to identify which elements of the images hold the most relevant information.

FROM SIMULATION TO REALITY

- Mainly a Deep Reinforcement Learning problem - Sim2Real
- Make sure that policies learnt in simulation are safely transferred to the real world

SOTA Techniques:

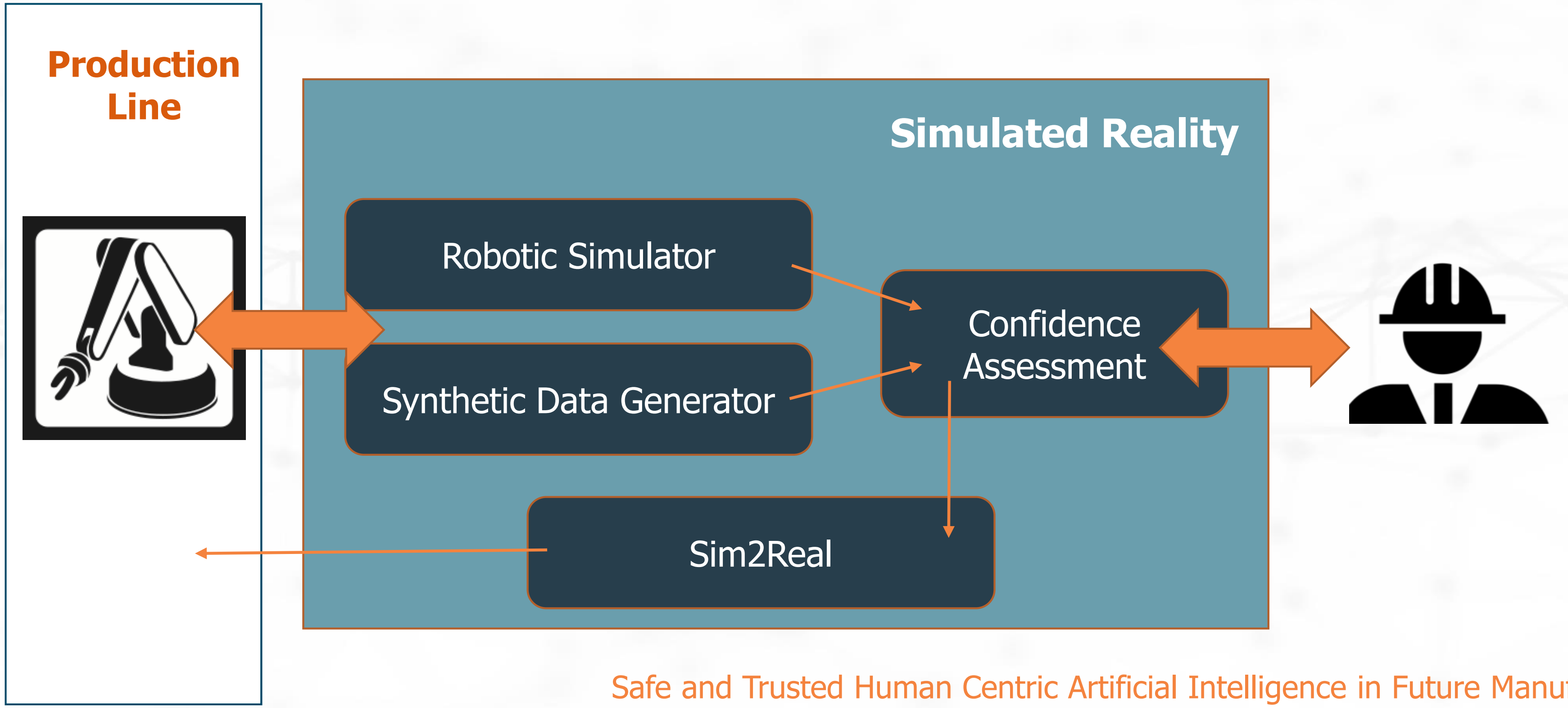
- Domain Adaptation – Shorter round of training in reality to adapt knowledge gained in simulation
- Domain Randomization – Produce different simulated training conditions with randomization
- Randomized-to-Canonical Adaptation Networks (RCANs) - Convert real world episodes to their simulated equivalent

UC3. RELIABLE DATA AUGMENTATION (1)

- Addresses the lack of sufficient training data and data skewness (e.g. defective parts much fewer than non-defective)
- In Supervised Learning (e.g. Visual Quality Inspection): Synthesis of training samples based on existing ones through:
  - Computer Vision (Rotation, Deformation, Noise etc.)
  - Generative Adversarial Networks
  - Variational Auto Encoders
- In Reinforcement Learning (e.g. Part Handling):
  - Imitation Learning through robot trajectory logs or human control
  - Reduces amount of trial and error to achieve the task

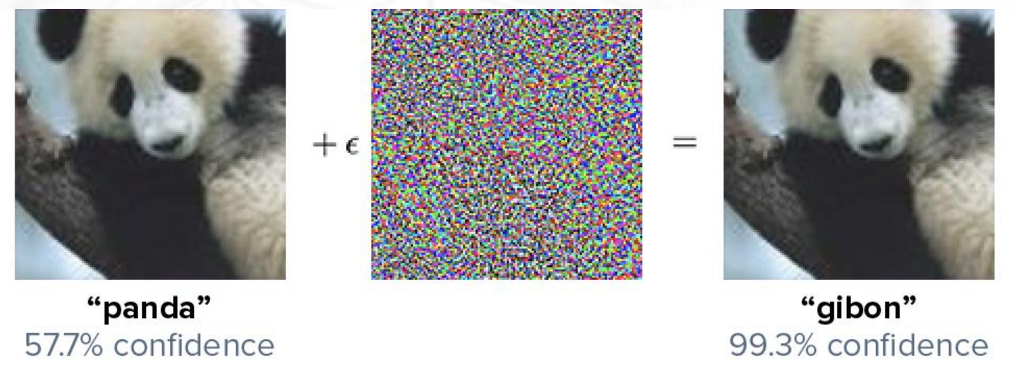
UC3. RELIABLE DATA AUGMENTATION (2)

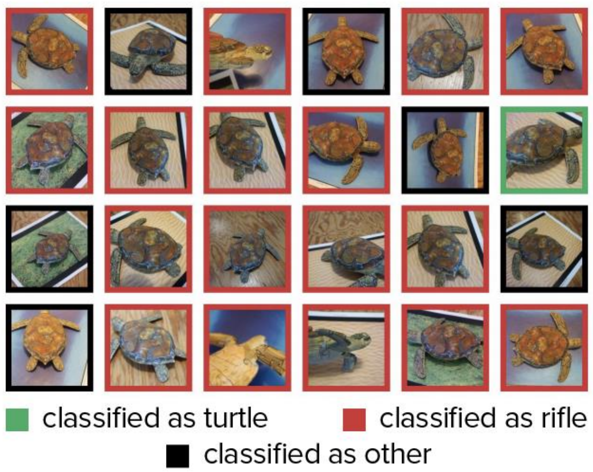
- Simulated Reality:
  - Mainly a Deep Reinforcement Learning problem - Sim2Real
  - Policies learnt in simulation are safely transferred to the real world



UC4. IDENTIFYING ADVERSARIAL ACTIONS AND CYBERSECURITY ATTACKS

- Security vulnerabilities coming from AI model errors have become a real concern - State-of-the-art deep neural networks can be easily fooled by a malicious actor and thus made to produce wrong predictions
- Two main pillars:
  - Explore strategies to generate adversarial examples
  - Explore Defenses Against Adversarial Examples
- Goal: Detection mechanism for pinpointing the adversarial examples



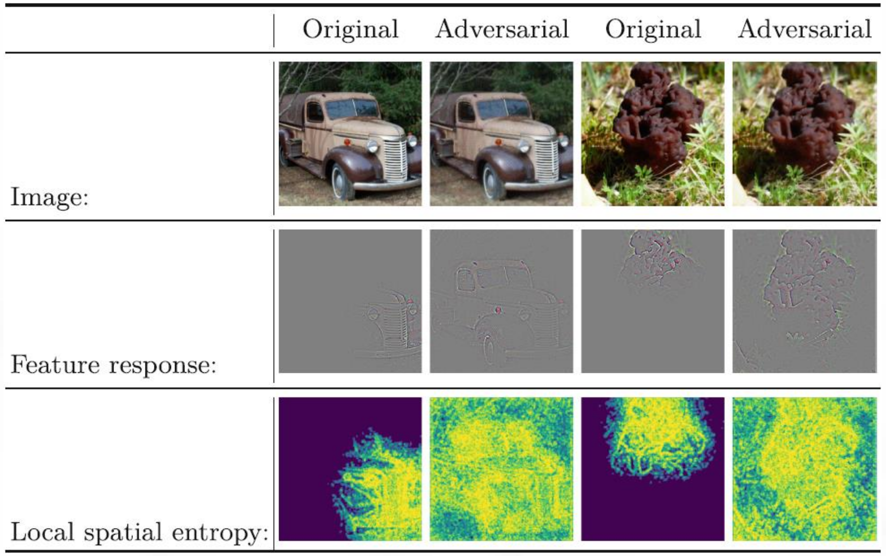


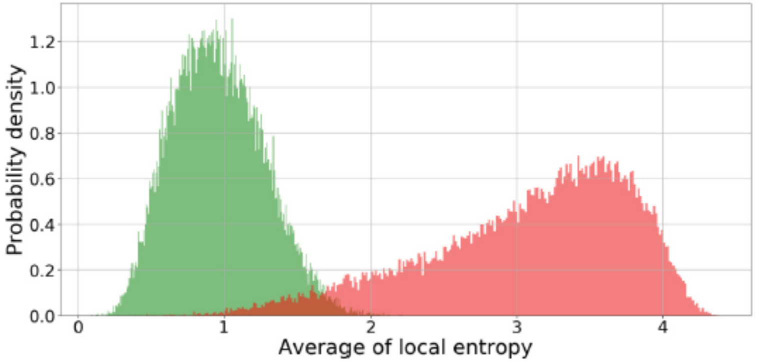
IDENTIFY ADVERSARIAL ATTACKS THROUGH FEATURE RESPONSE MAPS

- Model Specific Method (CNNs)
- Create adversarial attacks through novel methods (FGSM, Deep Fool, Grad Attack etc.)
- Create a feature response for given input
  - XAI Methods from Library (Guided Backpropagation etc.)
- Detect attacks based on the statistical analysis of spatial entropy

$$S_k = -\sum_i \sum_j \boldsymbol{h}_k(i,j)\log_2(h_k(i,j))$$

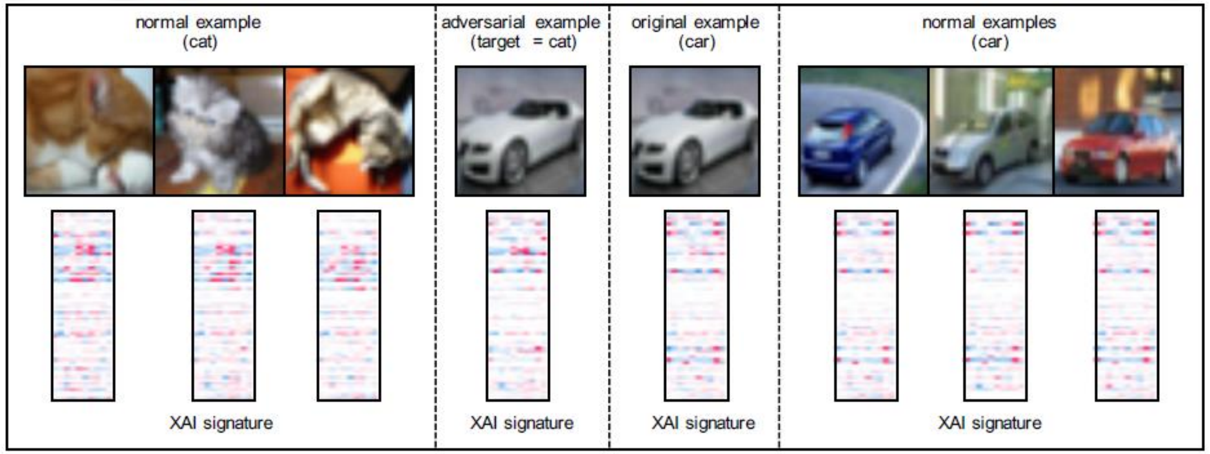
Other metrics can also be used such as Correlation Coefficient (CC) and Dice Similarity Coefficient (DSC)



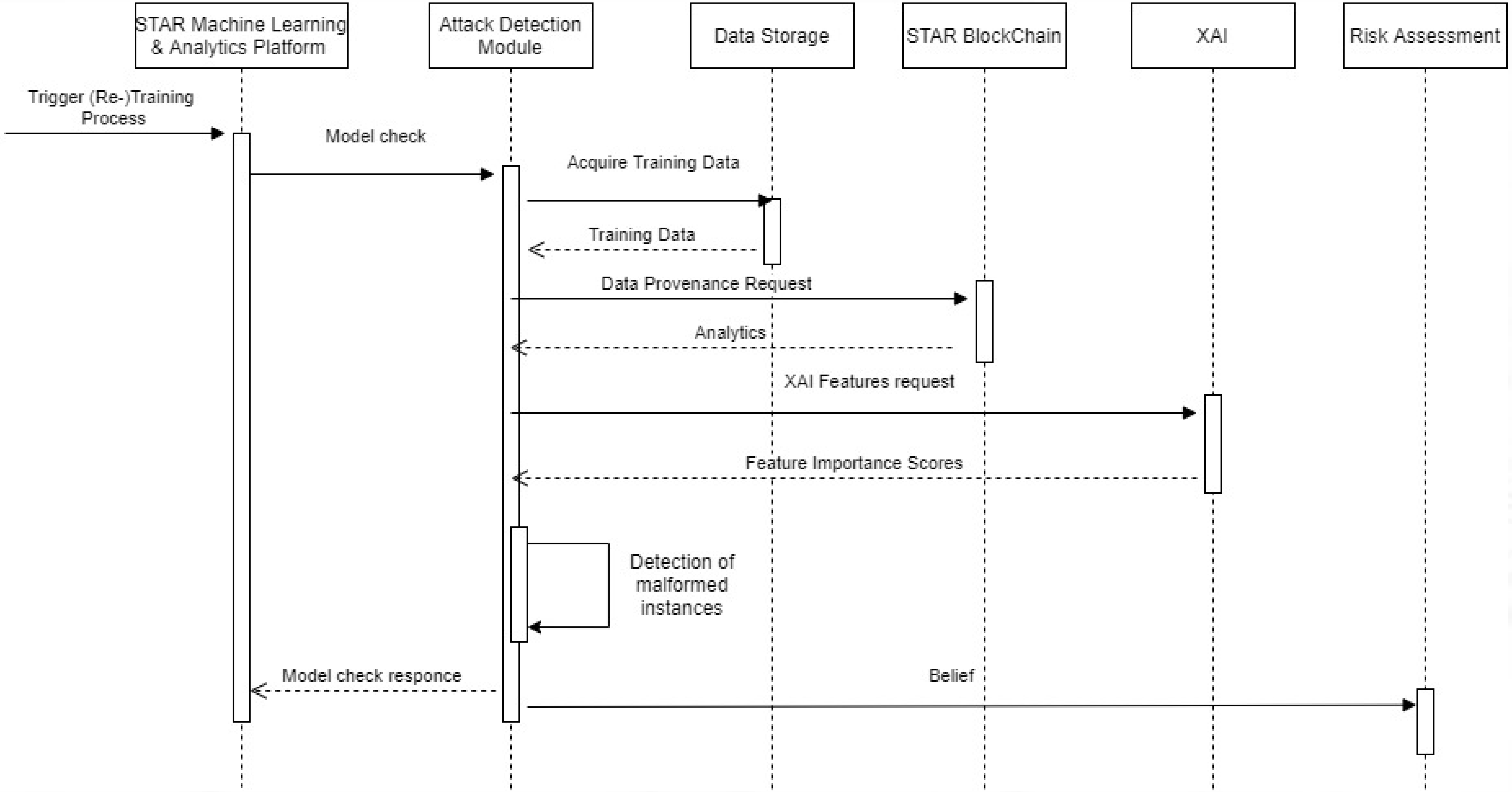


IDENTIFY ADVERSARIAL ATTACKS THROUGH SHAP SIGNATURES

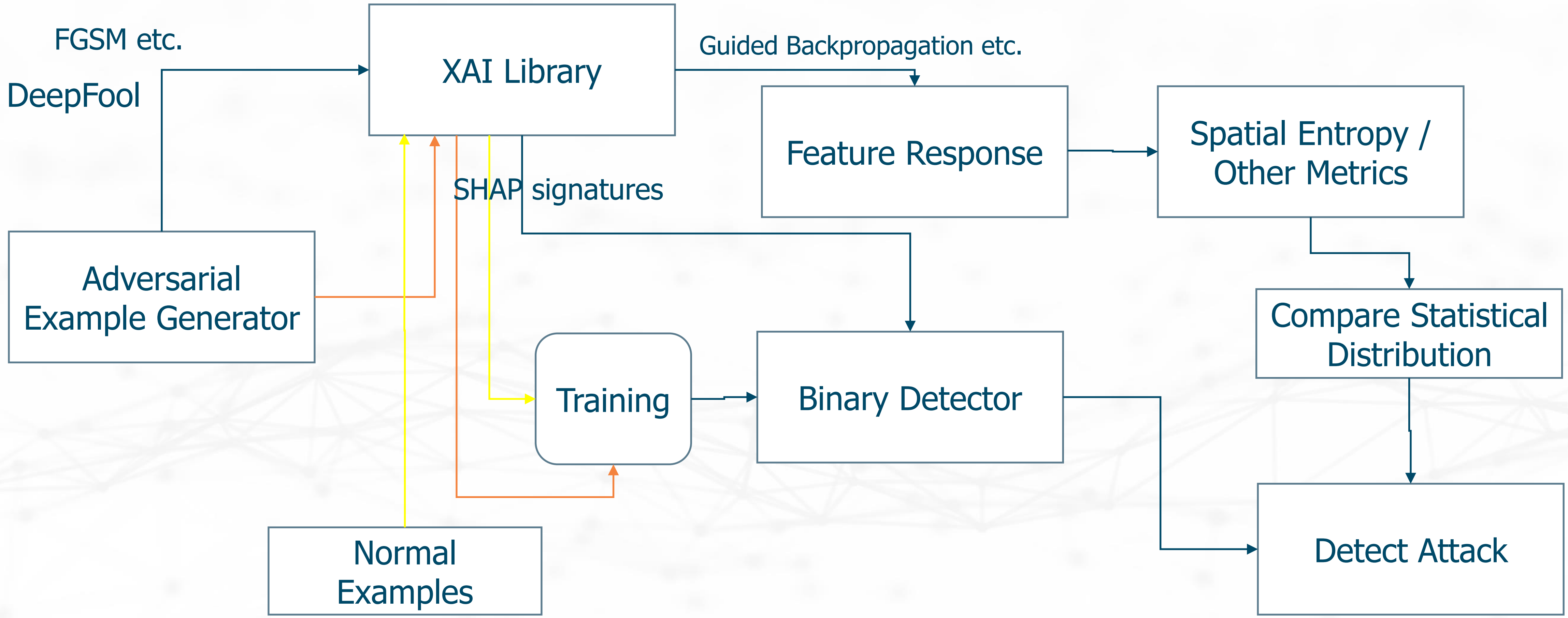
- XAI method: SHAP
  - Explain the model using Shapely values
  - Concept from game theory
  - Estimate the contribution of a specific input or neuron to a model decision
- Compute importance scores of the neurons of the penultimate layer of the classification model
  - Then use important scores as features for the detector
- Train a binary detector based on the SHAP signatures of adversarial/normal examples



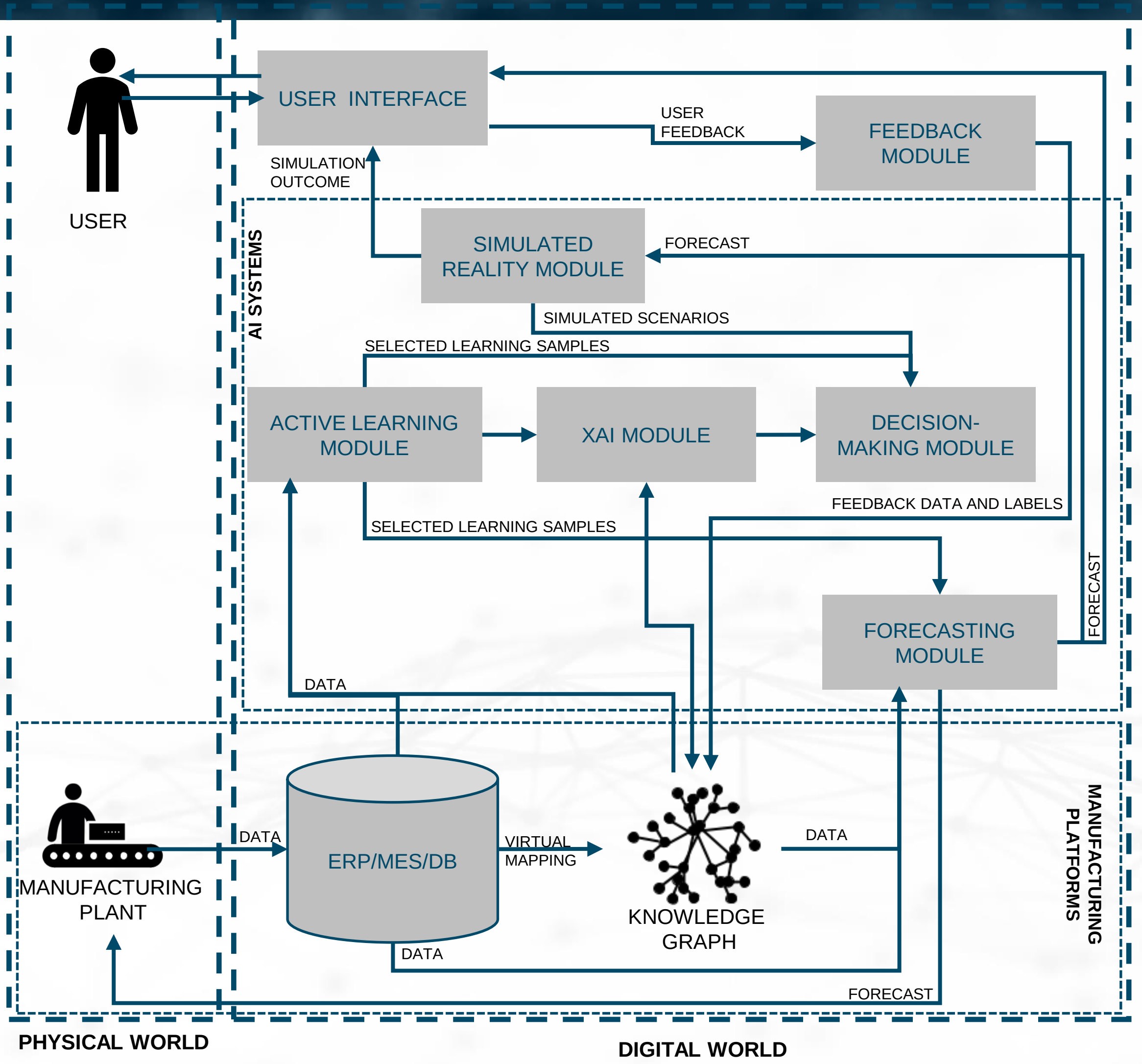
XAI FOR CYBERDEFENCE

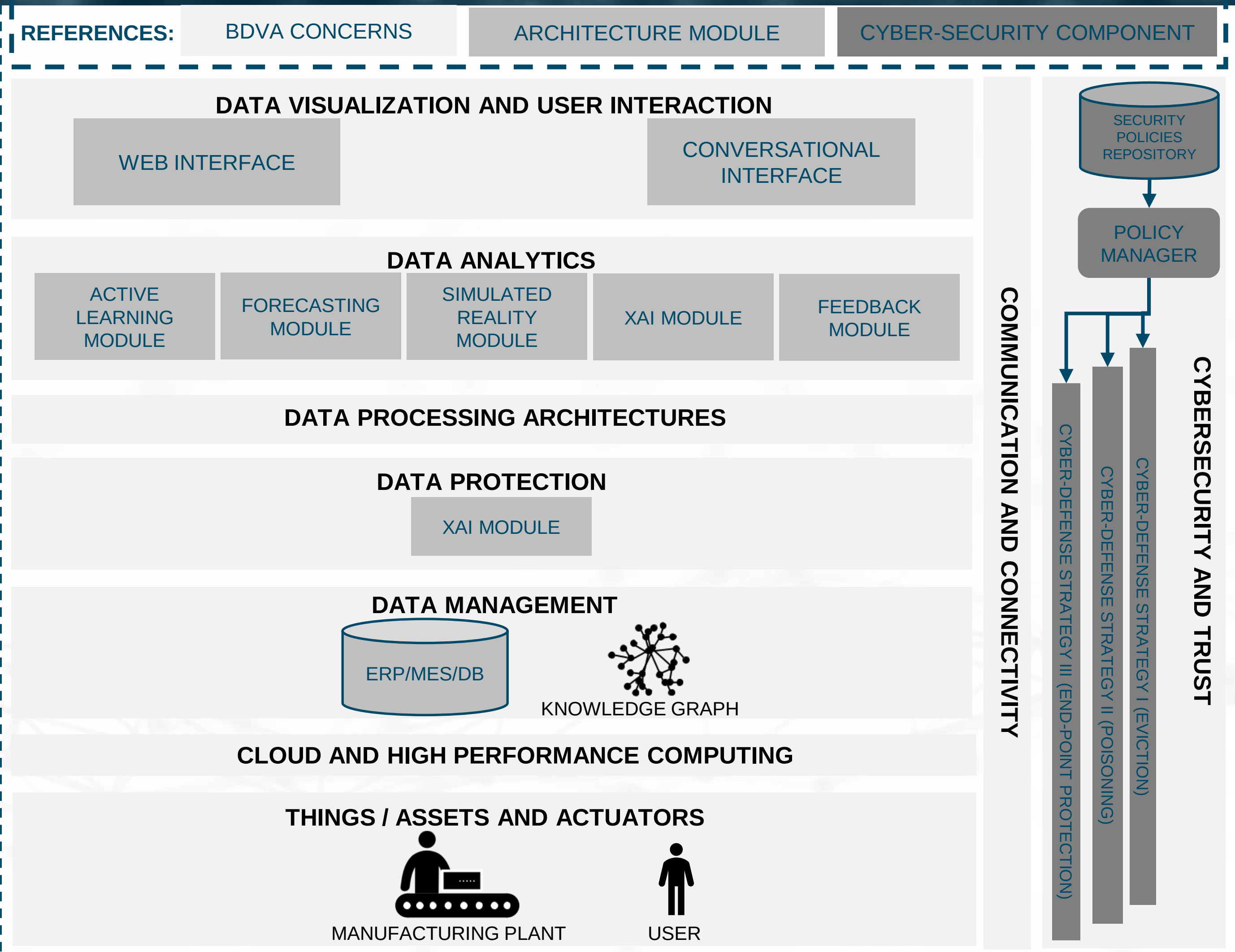


XAI-BASED SYSTEM ARCHITECTURE



XAI IN ACTIVE LEARNING





THANK YOU FOR YOUR ATTENTION

John Soldatos

INTRASOFT International

John.Soldatos@intrasoft-intl.com

www.star-ai.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°956573