IOTWeek

Dublin — June 20-23, 2022

Trusted IIoT & AI Technologies for the Future of Manufacturing

John Soldatos, Netcompany-Intrasoft

John.Soldatos@Netcompany-Intrasoft.com

GLOBAL VISION:

IoT TODAY AND BEYOND

netcompany

intrasoft

IOTForum

Trust in IIoT & AI Systems: The link with the AI Act

IOTWeek
Dublin — June 20-23, 2022

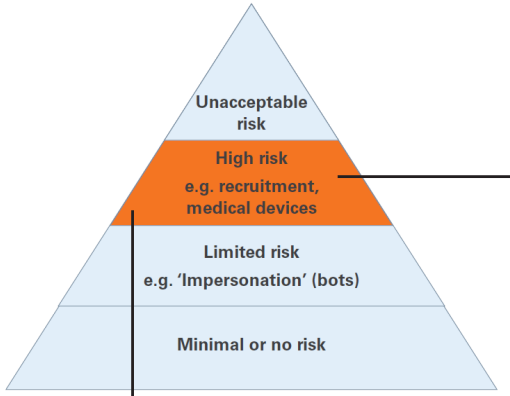
netcompany
intrasoft

High Quality Data for AI training, testing etc.

High Robustness & Cybersecurity

Explainability

Human Oversigth



REMOTE BIOMETRIC IDENTIFICATION (RBI)

Putting on the market of RBI systems (real time and ex-post)

- **Ex ante third party conformity assessment**
- Enhanced logging requirements
- ''Four eyes'' principle
  - No additional rules foreseen for the use of real-time and post RBI systems: existing data protection rules apply

1. **Safety components of regulated products** (e.g. medical devices, machinery) which are subject to third-party assessment under the relevant sectorial legislation
2. **Certain (stand-alone) AI systems in the following areas**
   - Biometric identification and categorization of natural persons
   - Management and operation of critical infrastructure
   - Education and vocational training
   - Employment and workers management, access to self-employment
   - Access to and enjoyment of essential private services and public services and benefits
   - Law enforcement
   - Migration, asylum and border control management
   - Administration of justice and democratic processes
3. Establish and implement **risk management** processes & in light of the **intended purpose** of the AI system
   - Use high quality **training, validation and testing data** (relevant, representative, etc.)
   - Establish **documentation** and design logging features (traceability and auditability)
   - Ensure appropriate certain degree of **transparency** and provide users with **information** (on how to use the system)
   - Ensure **human oversight** (measures built into the system and/or to be implemented by users)
   - Ensure **robustness, accuracy** and **cybersecurity**

H2020 STAR: A Project for Trusted Automation & Artificial Intelligence in Manufacturing

IOTWeek
Dublin — June 20-23, 2022

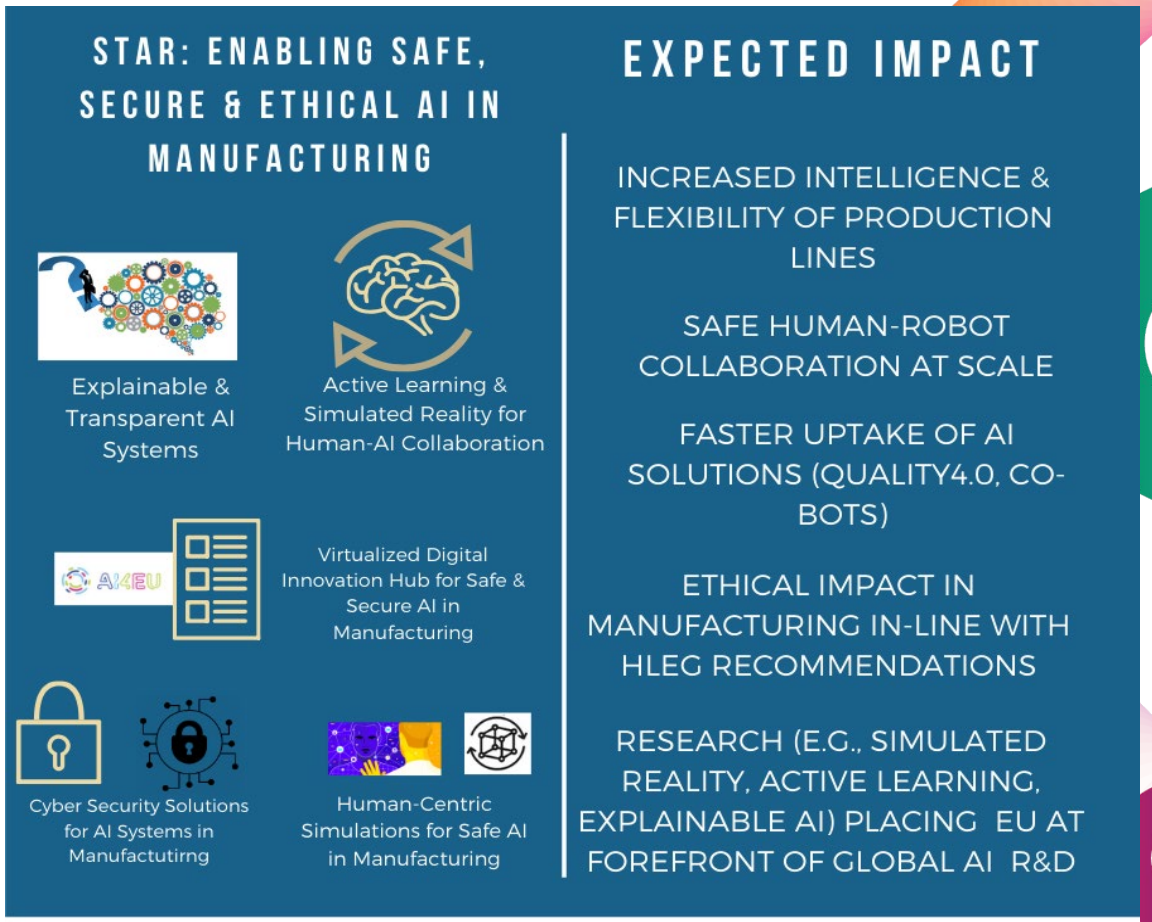
netcompany
intrasoft

Safe, Trusted and Human Centric AI in Manufacturing (Industry 5.0 Outlook)

STAR helps manufacturers and industrial automation vendors to build and deploy Safe Reliable and Trusted Human Centric AI systems

Enable AI systems to acquire knowledge in order to take timely and safe decisions in dynamic and unpredictable environments.

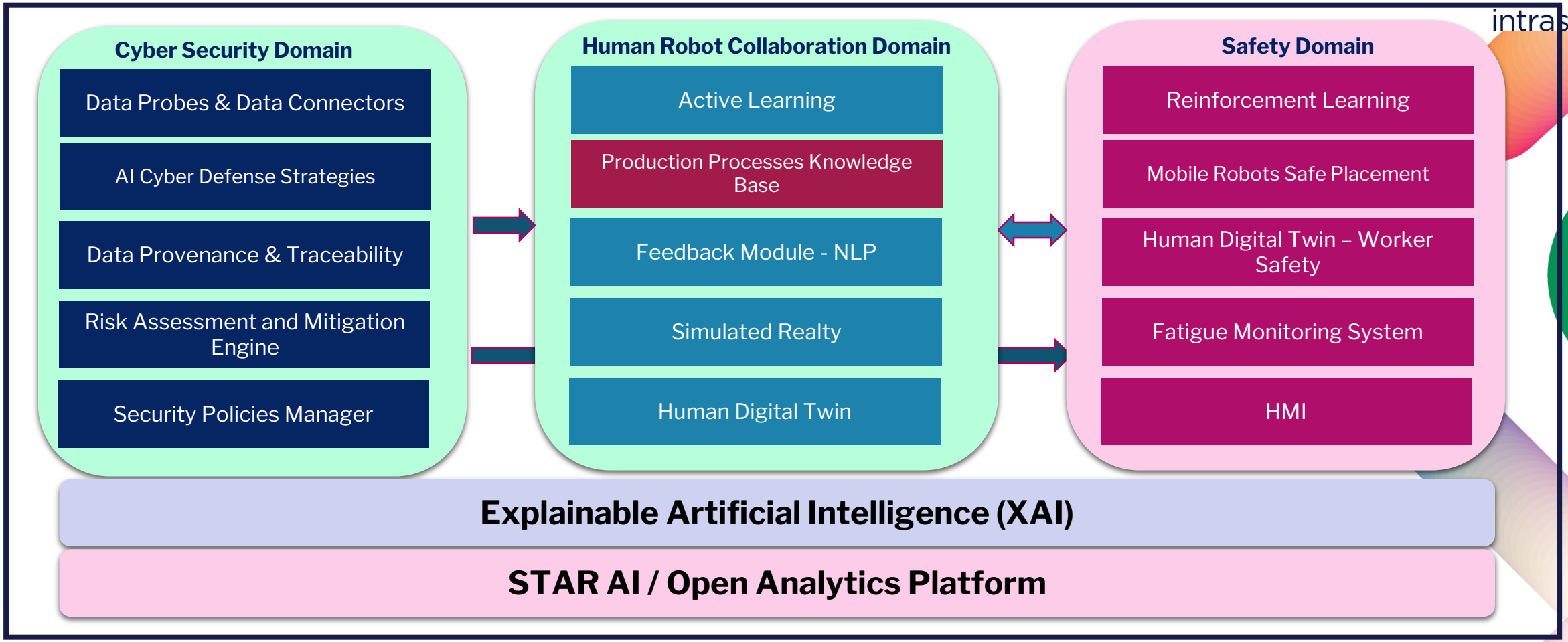
Build AI Manufacturing Systems in-line with Emerging Regulations (AIAct)



STAR Reference Architecture Model

IOTWeek Dublin June 20-23, 2022

netcompany intrasoft



Industrial Data Reliability

Environmental influences

- High or low temperatures, humidity, moisture, and air pressure factors)

Background noise

- Noise pollution, interference (alarms, extraneous speech), electrical noise (motors, cooling devices, air conditioning, power supplies)

Faulty or inaccurate sensors

- Sensing systems with poor precision.

Dying battery of a system

- Compromises its ability to operate properly and provide reliable measurements.

Compromised or attacked devices

- Produce biased or fake data due to adversarial attacks (e.g., data modification, false information injection).

Compromised AI or BigData analytics algorithms

- Algorithms under poisoning or evasion attacks

Data Reliability and attacks against AI System

netcompany

intrasoft

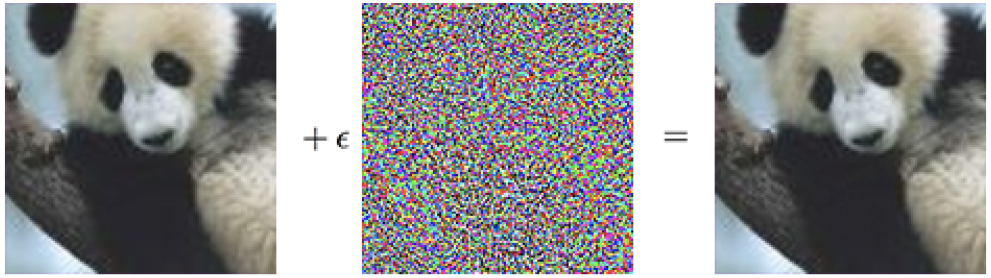
Security vulnerabilities coming from AI model errors have become a real concern

State-of-the-art deep neural networks can be easily fooled by a malicious actor and thus made to produce wrong predictions

Explore strategies to generate adversarial examples

Explore Defences Against Adversarial Examples

Goal: Detection mechanism for pinpointing the adversarial examples leveraging Explainable AI



“panda”
57.7% confidence

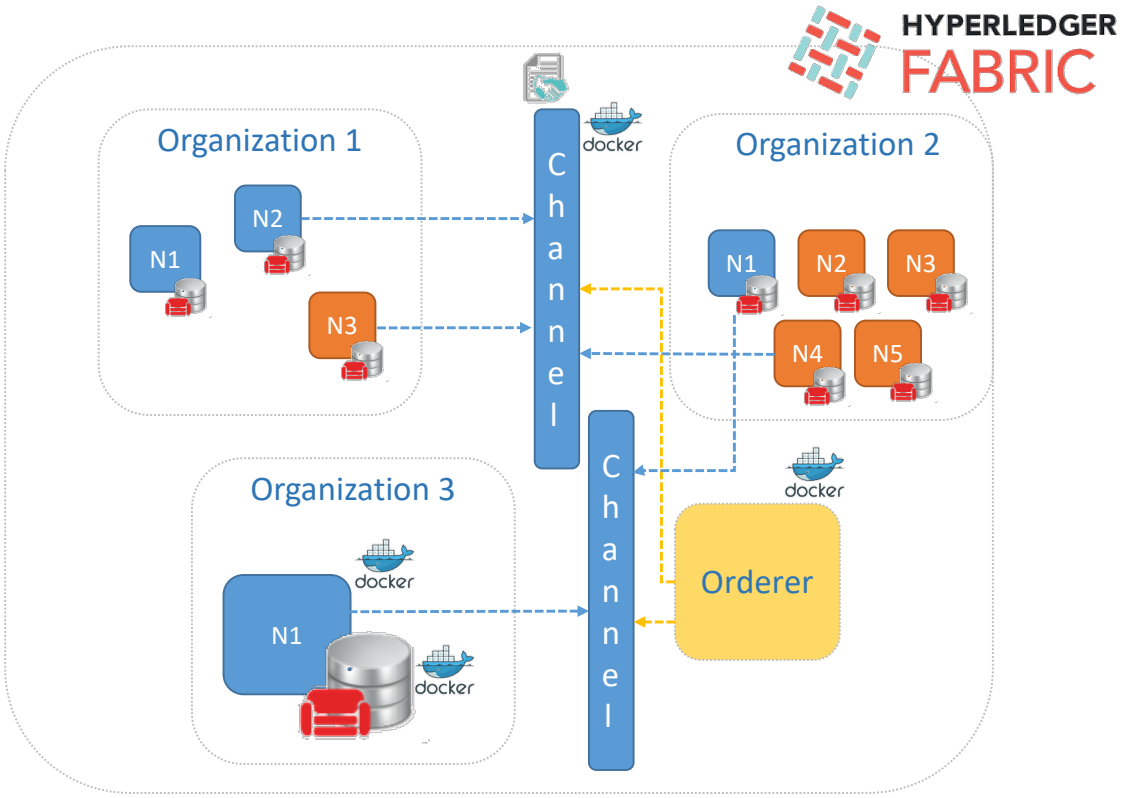
“gibon”
99.3% confidence

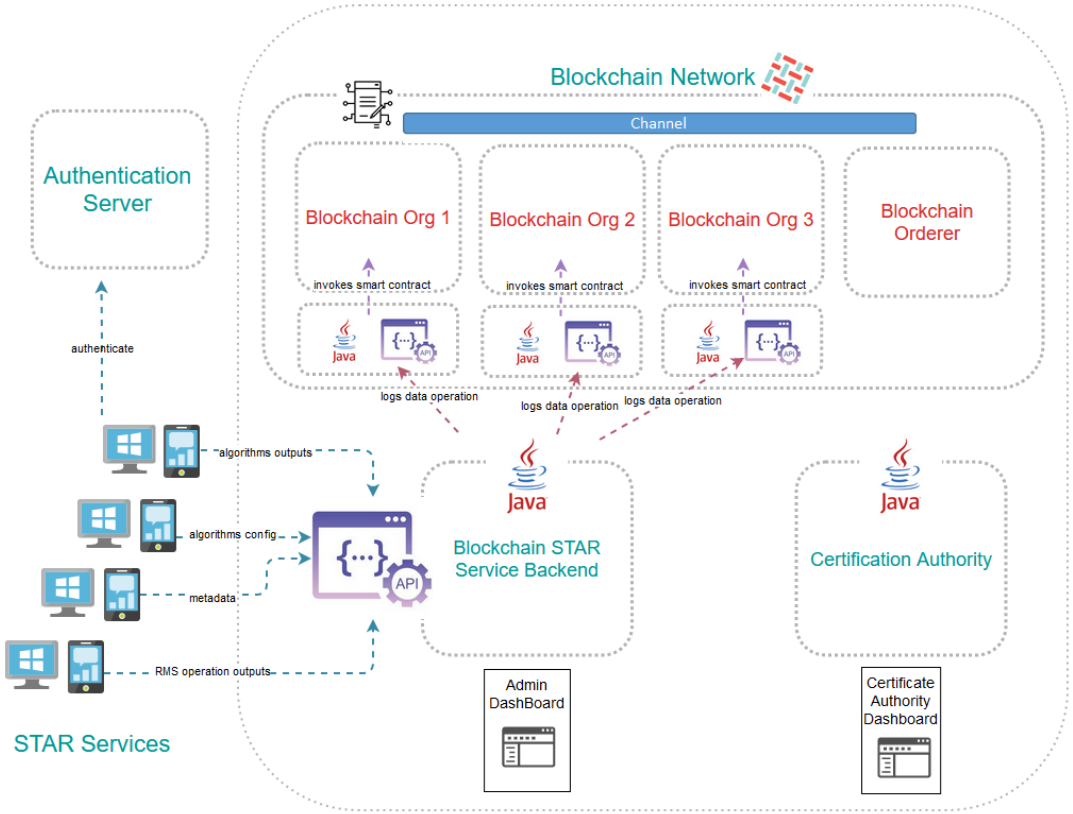
classified as turtle
classified as rifle
classified as other

STAR Data Provenance and Reliability Service

IOTWeek
Dublin — June 20-23, 2022

netcompany
intrasoft





STAR Blockchain Value Propositions

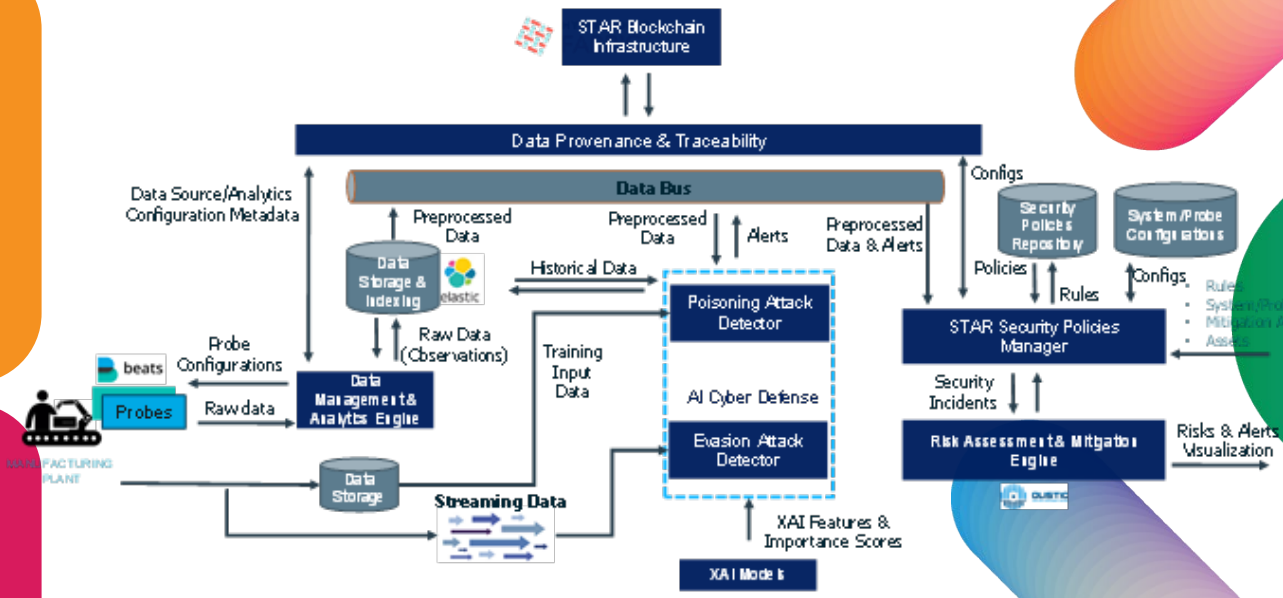
IOTWeek Dublin June 20-23, 2022

netcompany

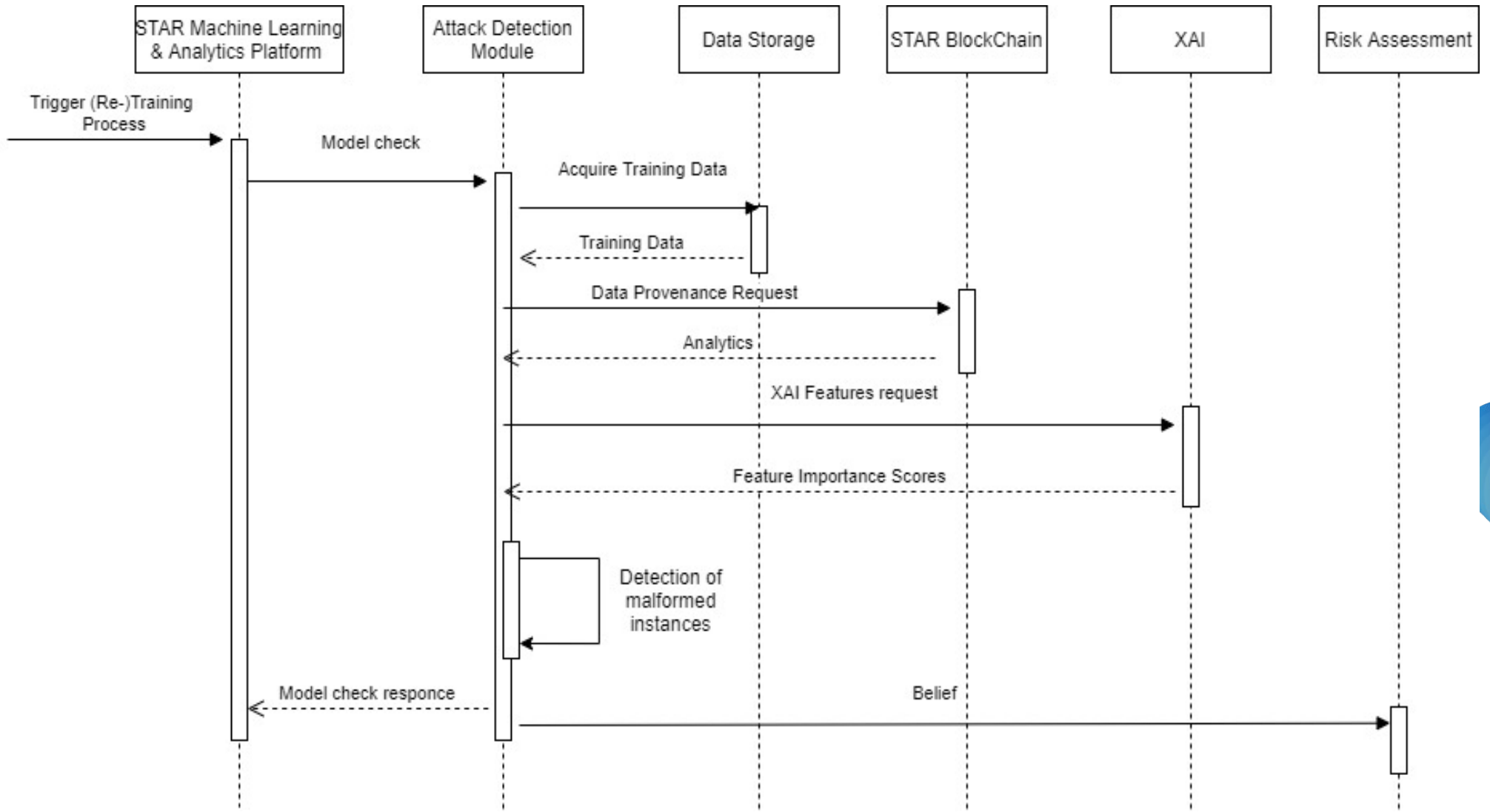
intrasoft

Reliable Algorithms: Provenance of AI algorithms metadata and configurations – “Sealed” algorithms

Reliable AI Outcomes: Provenance of AI analytics outcomes – “Sealed Outcomes”



Defending a Poisoning Attack



From Black-Box AI to Interpretable Models

netcompany
intrasoft

- 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

Uses of Explainable AI

netcompany
intrasoft

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.
- EU AI Regulatory Compliance

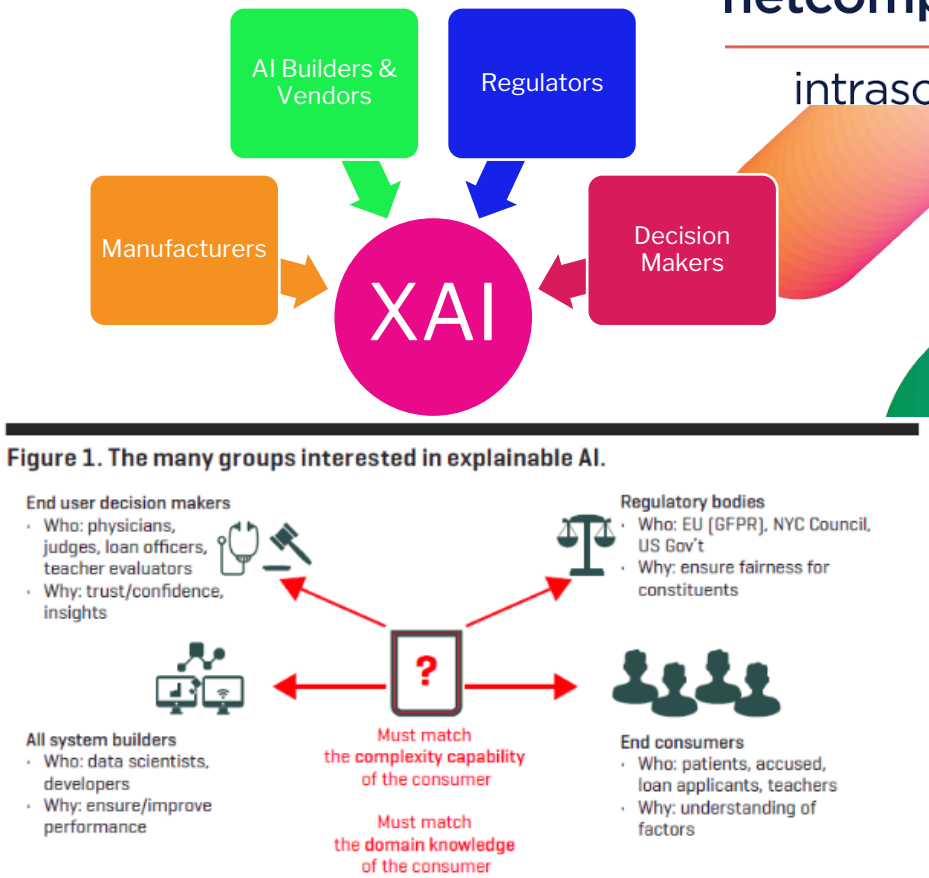


Figure 1. The many groups interested in explainable AI.

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

Explaining Quality Inspection – Why is a part defected?

IOTWeek
Dublin — June 20-23, 2022

netcompany

Explanations of classification models

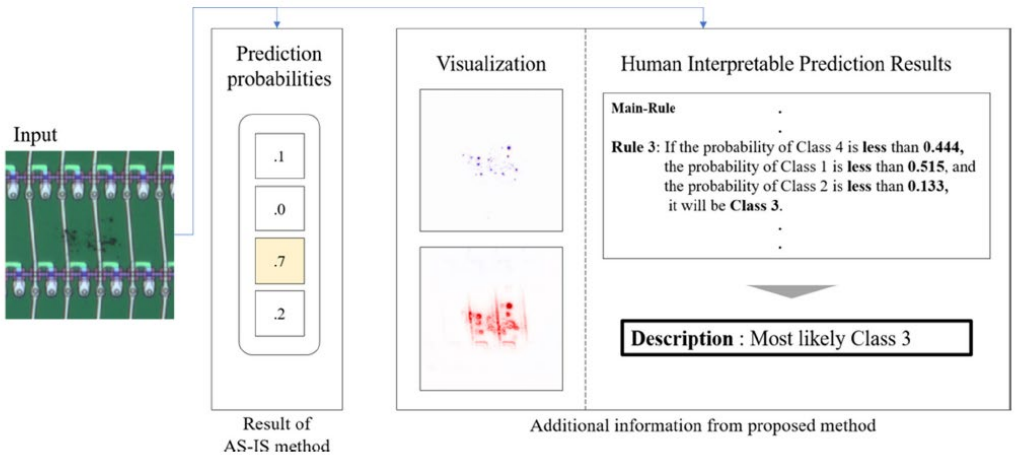
- Image data + Attribution methods
- Produce attribution maps + Visualize into heatmaps
- Highlight features responsible for or against the predicted class

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

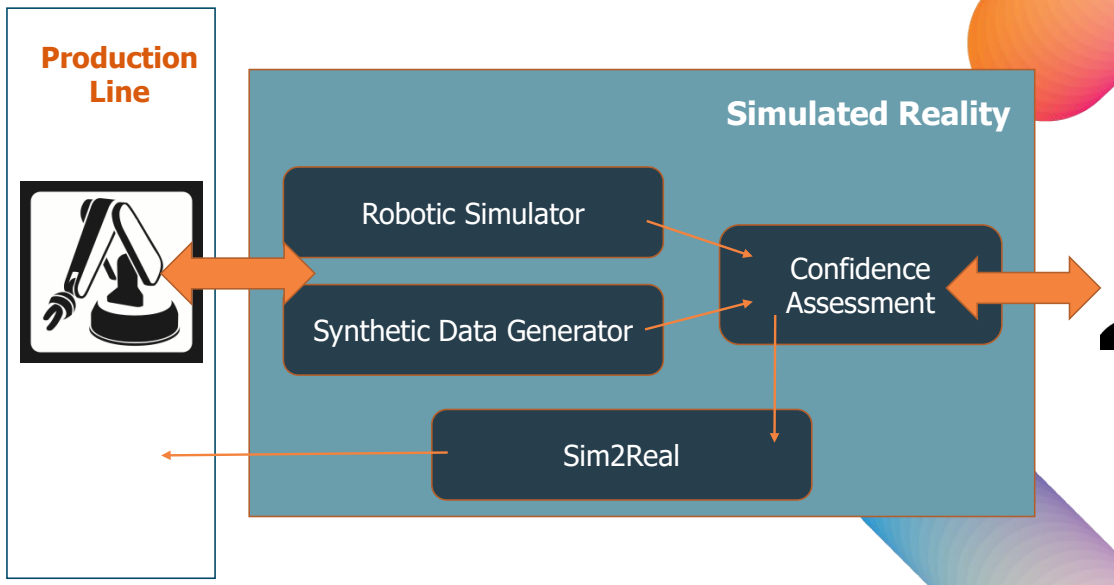


Simulated Reality

netcompany

intrasoft

- Policies learnt in simulation are safely transferred to the real world
- 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
- **Reliable Data Augmentation**: Addresses the lack of sufficient training data and data skewness (e.g. defective parts much fewer than non-defective)
- **Supervised Learning (e.g. Visual Quality Inspection)**: Synthesis of training samples based on existing ones :
  - Computer Vision (Rotation, Deformation, Noise etc.)
  - Generative Adversarial Networks
  - Variational Auto Encoders
- **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



More Information and Free Download

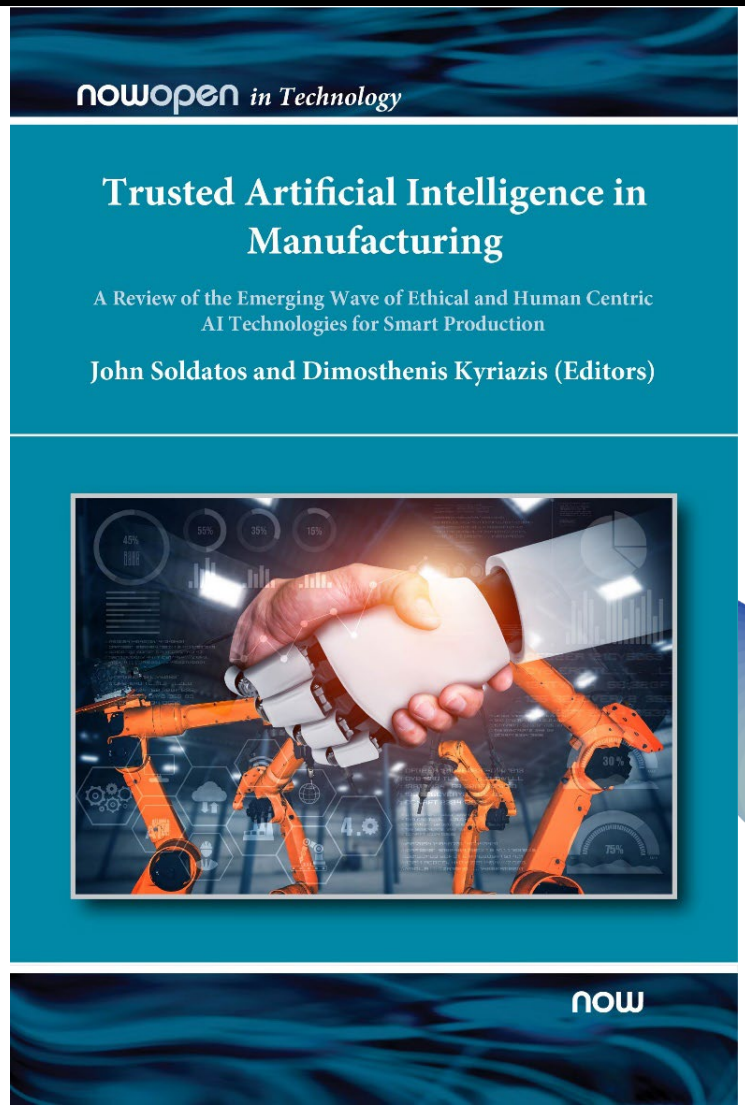
IOTWeek
Dublin — June 20-23, 2022

STAR Web Site: www.star-ai.eu

- Rich Library of Blogs and other publications

Open Access Book (published November 2021; 15500+ Downloads):

- John Soldatos (ed.), Dimosthenis Kyriazis (ed.) (2021), "Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production", Boston-Delft: now publishers,



netcompany
intrasoft

IOTWeek

Dublin — June 20-23, 2022

Thank you!

Find more:

STAR: https://www.star-ai.eu/

John Soldatos: https://www.linkedin.com/in/johnsoldatos/

iotweek.org