School of Engineering

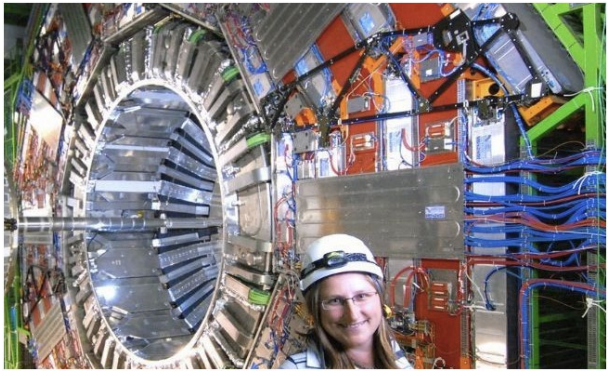
Trusting Artificial Intelligence in Safety Critical Systems - towards a comprehensive verification framework

Dr. Joanna Weng
ZHAW, School of Engineering (SoE)
Winterthur, Switzerland

1. Introduction

About Me

School of Engineering



10 years in research in Particle Physics@CERN, Geneva Switzerland

- Commissioning & monitoring of CMS detector
- Data analysis of first LHC collider data
- Discovery of Higgs Boson

Safety analysis & project management @NPP Mühleberg

- Probabilistic Safety Analysis
- First Swiss decommissioning safety analysis
- Presenting results to Swiss regulator

Zurich University of Applied Sciences (ZHAW)

School of Engineering

Senior Lecturer (Mathematics & Physics) & Applied Research:

- Safety-Critical Systems (SKS) group
- Associate Faculty Centre of AI (CAI)
  close collaboration with AI experts
- **TÜV Certified Functional Safety Engineer**
- **IEEE CertifAIEd Lead Assessor**

IEEE SA

TÜVRheinland®
Precisely Right.

Mission: Applied research projects in safety critical systems together with industry partner, research facilities or governmental agencies.

School of Engineering
CAI Centre for Artificial Intelligence

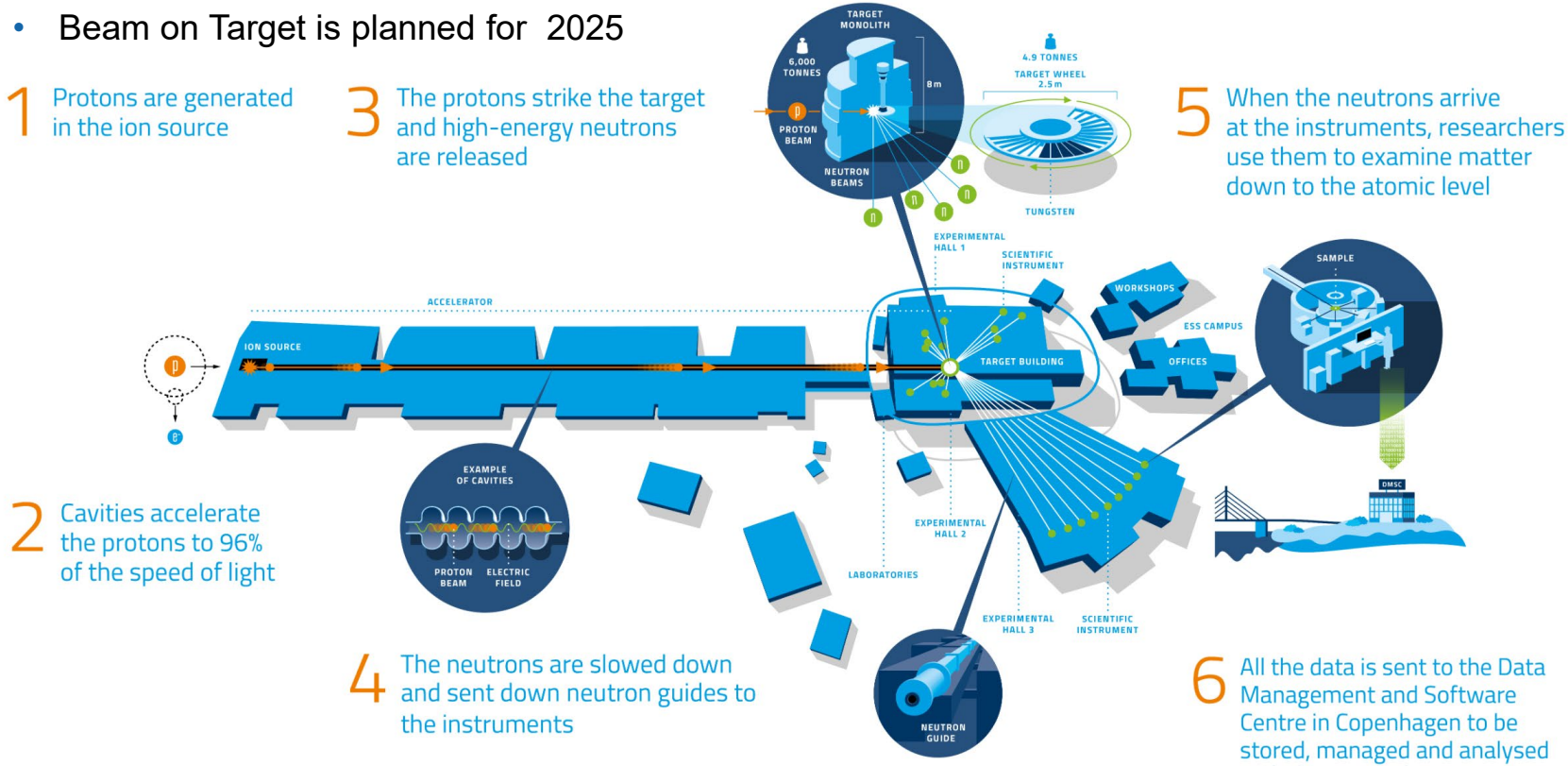
Research Funding: Directly by industry partners or national/international funding agencies

2. Example Project: ESS

European Spallation Source (ESS)

School of Engineering
IAMP Institute of Applied Mathematics and Physics

- ESS (Lund, Sweden) will be the brightest Neutron Source worldwide
- Beam on Target is planned for 2025

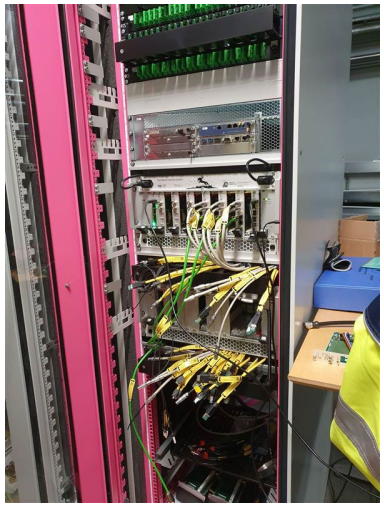


ESS Project

School of Engineering
IAMP Institute of Applied Mathematics and Physics

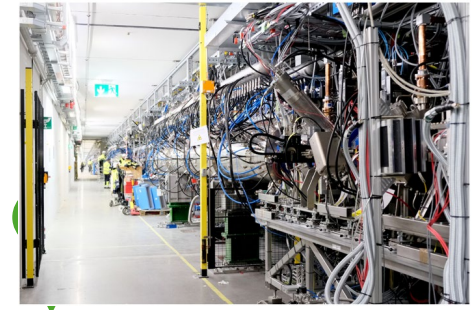
Successful collaboration with ESS since 2015:
(https://europeanspallationsource.se/)

- Designed, build and tested a distributed Fast Beam Interlock System (**FBIS**) in our lab at ZHAW
- Developed verification for FBIS modules (Hardware in the Loop (HIL) simulation)
- Performed FBIS Hardware Integrity assessment (Functional Safety Standards IEC61508, IEC61511)
- Support in Machine Protection System (MPS) Design and Reliability Assessment (FTA, RBD, ET)
- Design and Functional Safety assessment of ESS Personal Safety System (PSS) as externa assessor
- Support towards Swedish regulator in Licencing Process



FBIS

School of Engineering
IAMP Institute of Applied Mathematics and Physics

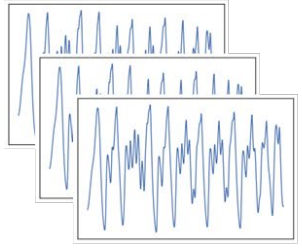


Pilot project APBIS: Autonomous Predictive Beam Interlock System@ESS

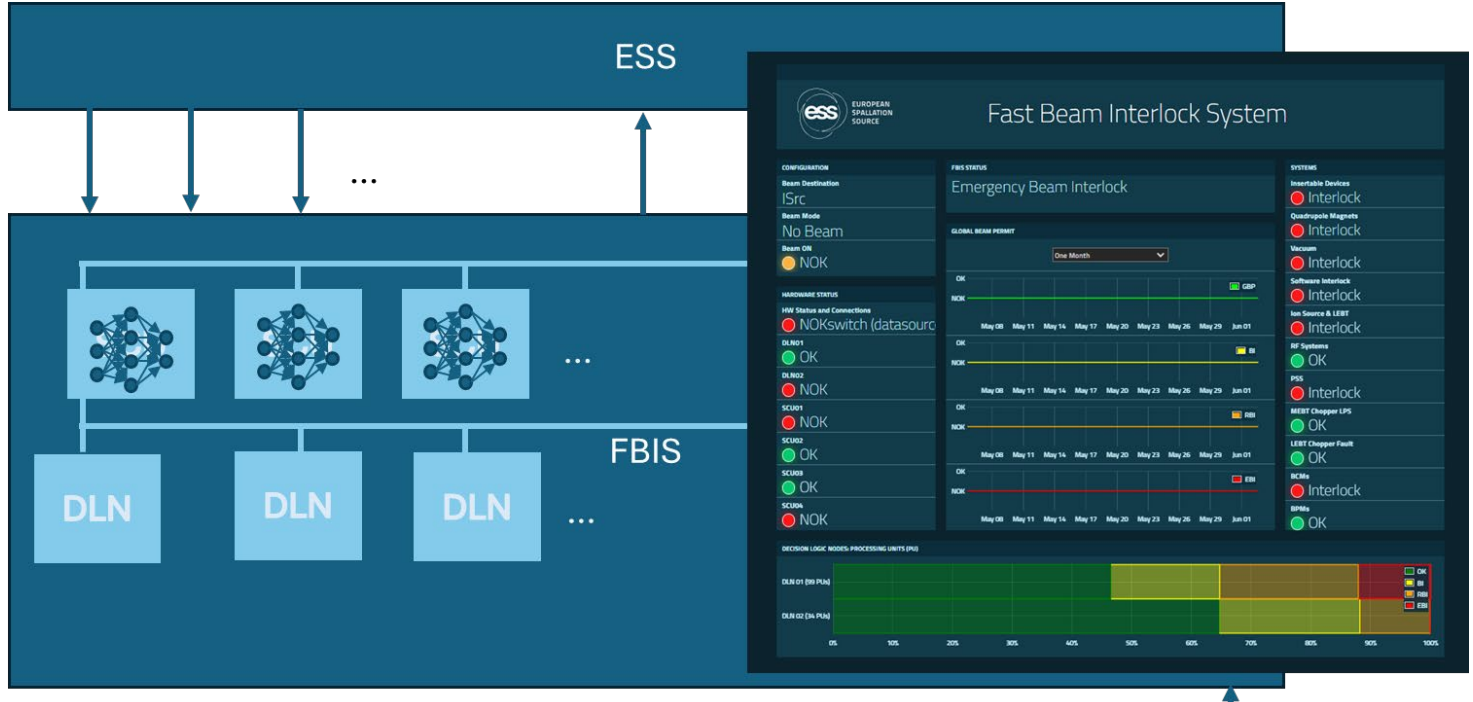
School of Engineering
IAMP Institute of Applied Mathematics and Physics

- Machine-learning model capable of analyzing sequential data streams in real-time
- ML-model is implemented on existing FBIS hardware (FPGAs)
- Assist FBIS in predicting the beam behavior
- Detect unwanted system conditions for further analysis purposes (anomaly detection)

ESS Sensor Inputs



APBIS



Additional Sensor Input to FBIS from APBIS

School of Engineering
IAMP Institute of Applied Mathematics and Physics

3. Research on AI in in Safety Critical Systems: Verification& Certification

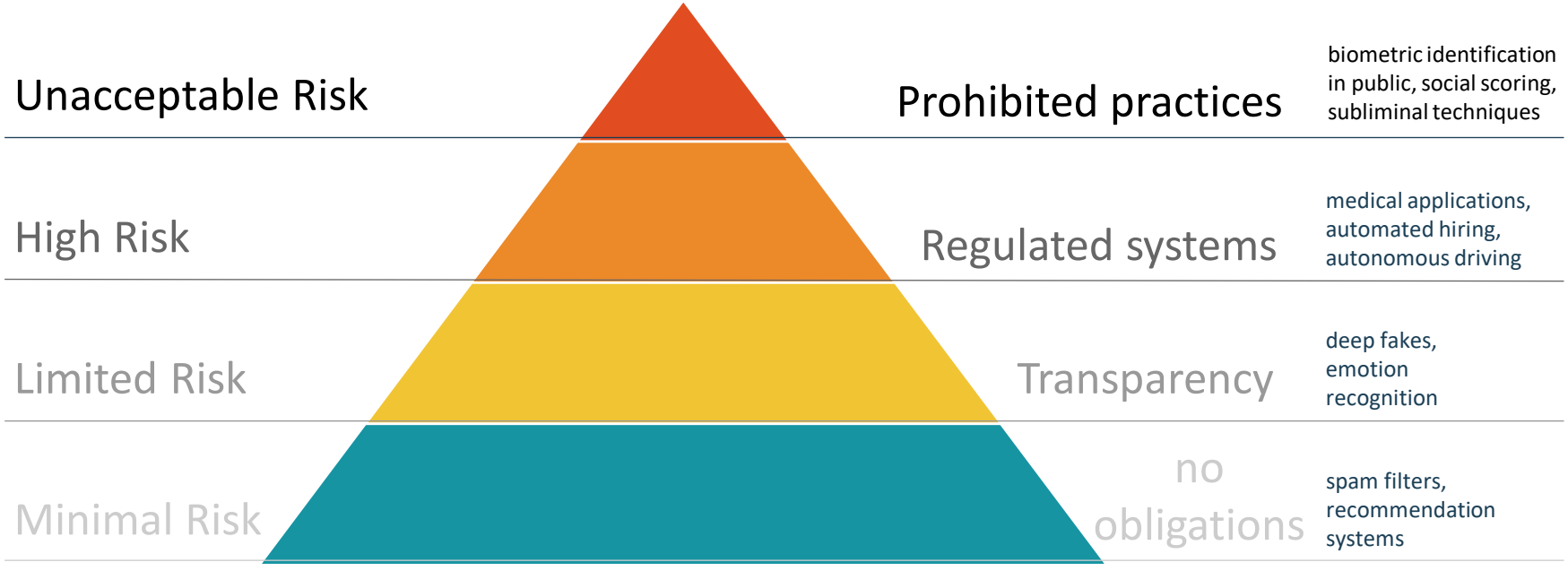
The EU AI Act

- After final approval by the Council of the EU on May 21, 2024, the EU AI Act is now set to be published in the EU's Official Journal.
- The EU AI Act will enter into force on the 20th day after publication
- Companies will have **2-3 years** to adapt to the regulation if they want to access the EU market

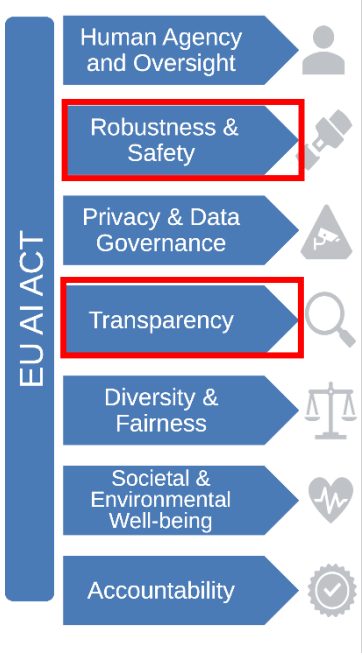
EU AI Act follows risk-based approach:

High risk systems (e.g. AI systems in critical infrastructure) will be **regulated**.

EU AI Act: Risk-based approach



EU AI Act: Dimensions of trustworthy AI



What degree of **autonomy** is appropriate?

Is the behaviour of the AI component **consistent** and **functionally safe**?
How does it hold up against **attacks**?

Do the (training and input) data protect **privacy** and **company secrets**?

Are the AI functions and decisions made by the AI **comprehensible**?

Are **minorities** fairly treated?

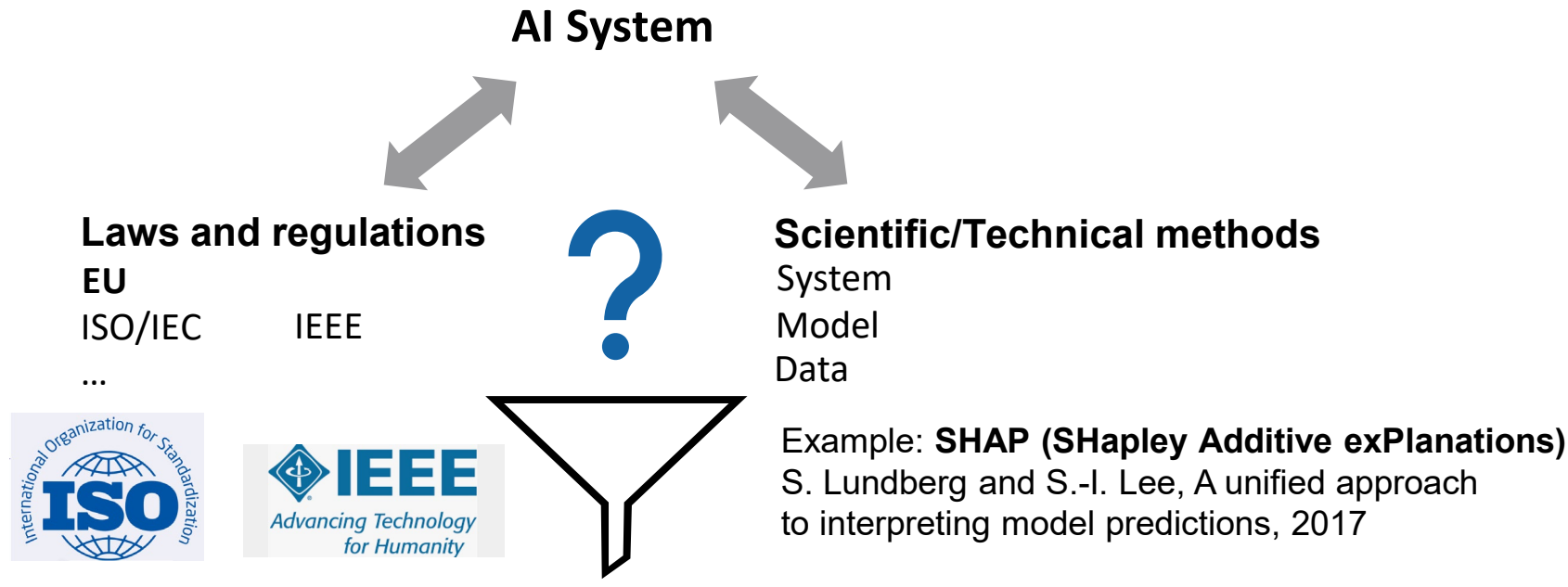
What is the **impact** on society and the environment?

Who is responsible for the correct **development**, **deployment**, and **operation**?

Bridging the Gap

There is a s gap between current (developing) **regulations** for AI trustworthiness and **concrete guidelines** including **scientific methods.**

Our research aims to bridge this gap.

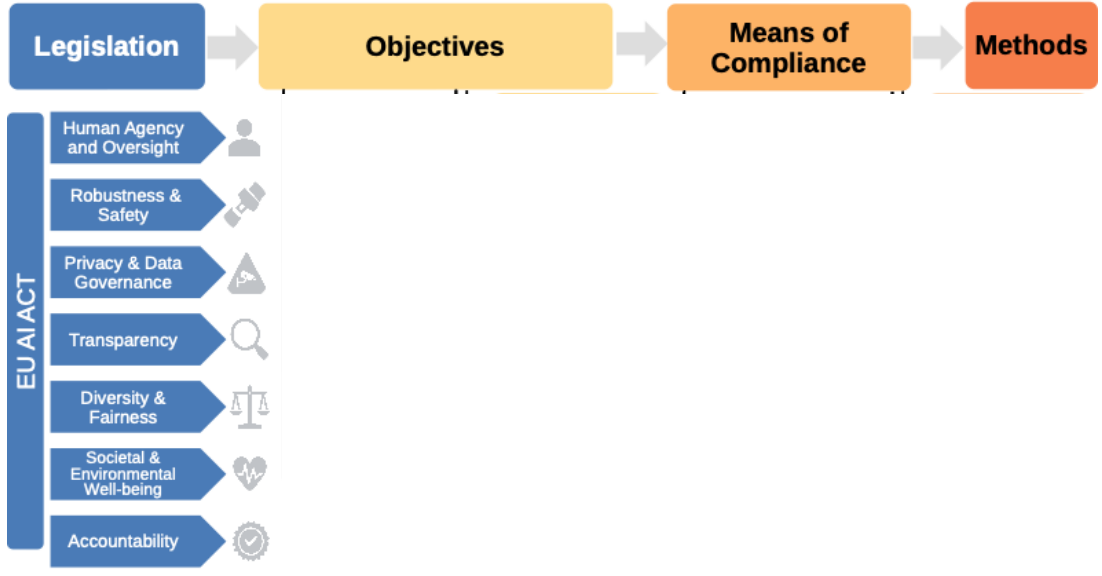


Project certAInty

School of Engineering

Development of a **certification scheme** for AI-based systems (together with Swiss certification company CertX as research partner)

- Dimensions for trustworthy AI in line with the EU AI act



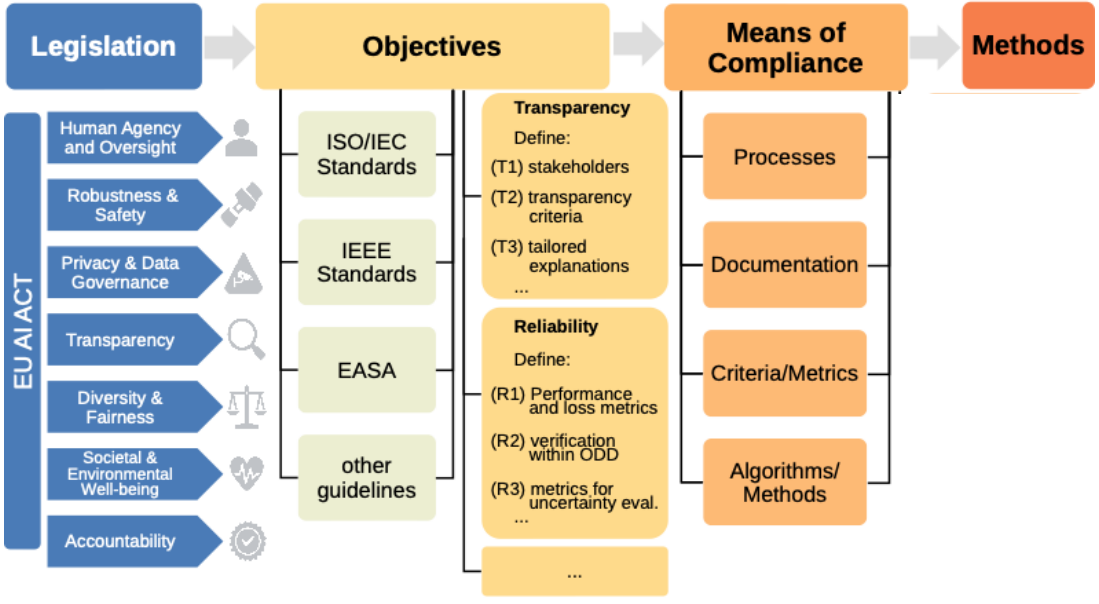
Innovation project supported by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Innosuisse – Swiss Innovation Agency

Project certAInty

School of Engineering

Development of a **certification scheme** for AI-based systems

- Objectives and Means of Compliance derived from these legal obligations and standards



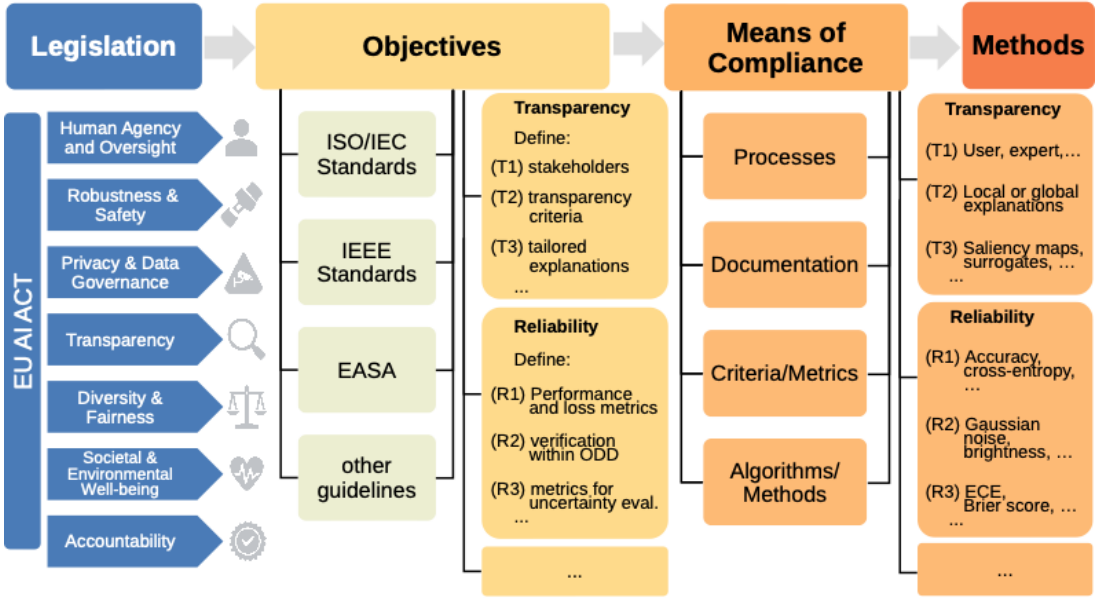
Innovation project supported by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Innosuisse – Swiss Innovation Agency

Project certAInty

School of Engineering

Development of a certification scheme for AI-based systems

- Link between objectives and technical methods



Innovation project supported by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Innosuisse – Swiss Innovation Agency

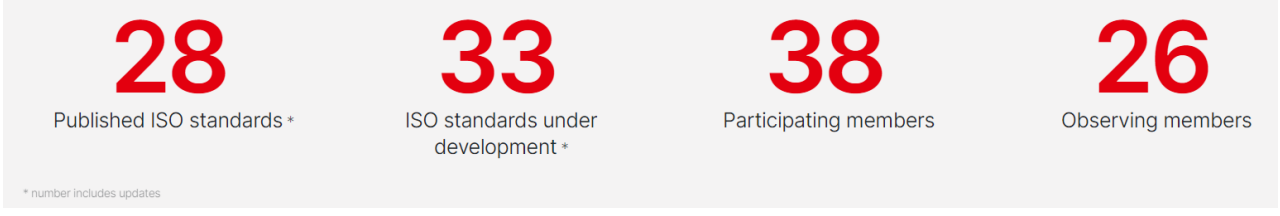
Example standards: ISO/IEC

The International Electrotechnical Commission (IEC) and the International Organization for Standardization (ISO) are global organizations that develop and publish international standards for a wide range of industries.

Focus

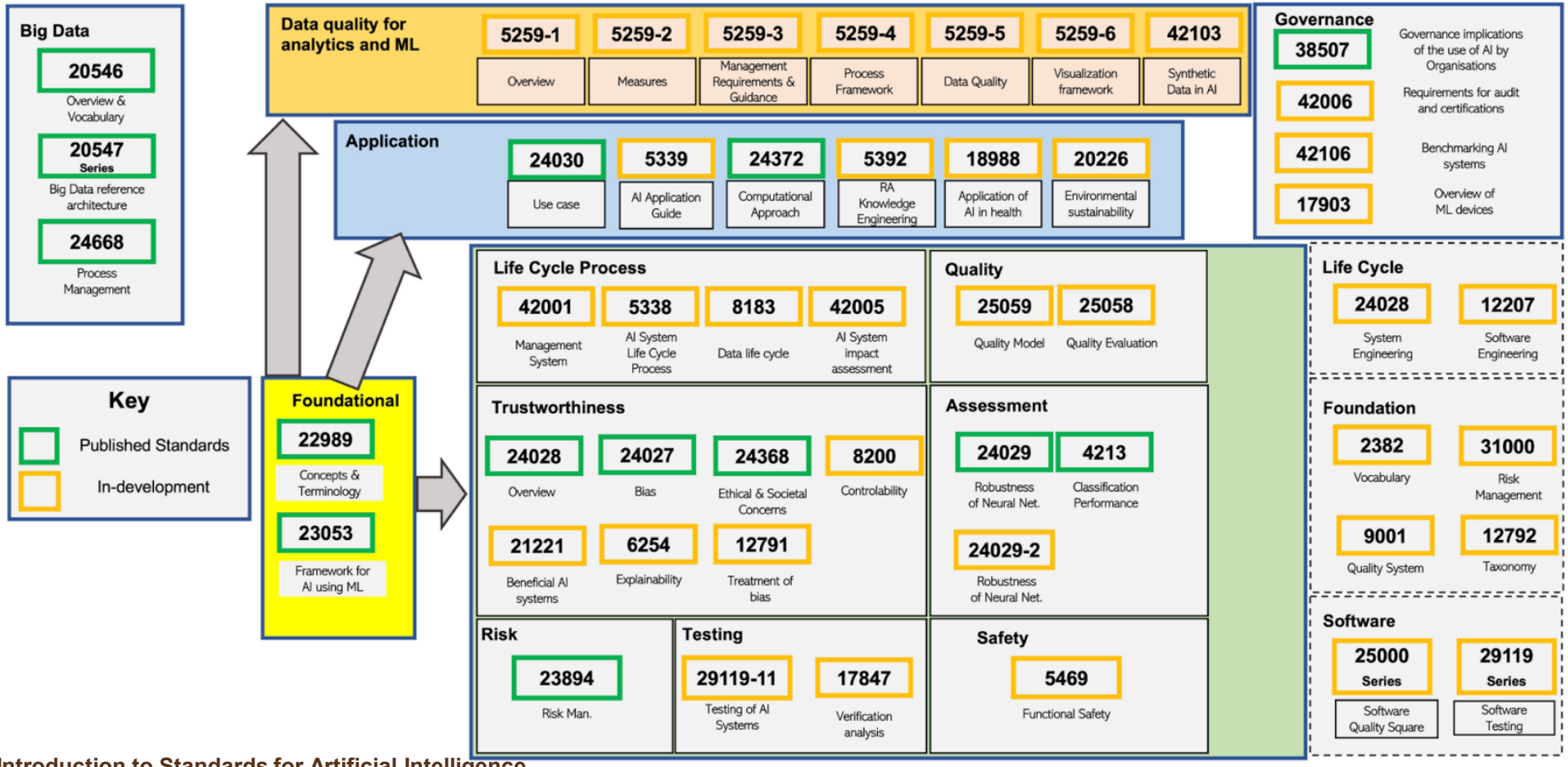
Sociotechnical systems
Technical standards

Topics of EU AI act covered (JTC 1/SC 42)



ISO/IEC

School of Engineering
IAMP Institute of Applied Mathematics and Physics



Ref.: Introduction to Standards for Artificial Intelligence

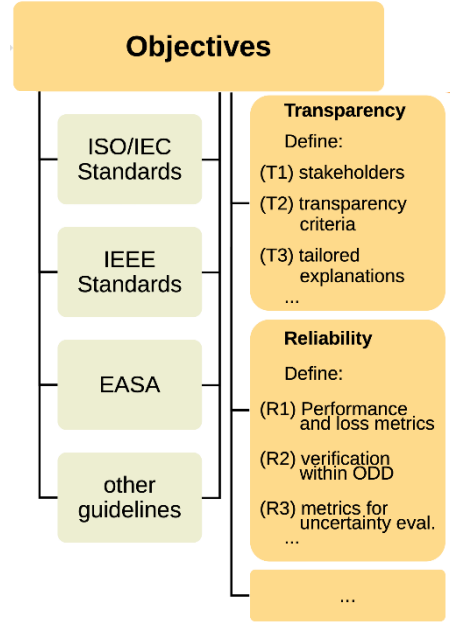
Certification Scheme: Objectives

Objectives based on different **standards and guidelines**
(38 documents in total)

- ISO/IEC
- IEEE
- EASA
- Fraunhofer Institute

Objectives specified for **specific** trustworthiness **dimensions**, e.g.

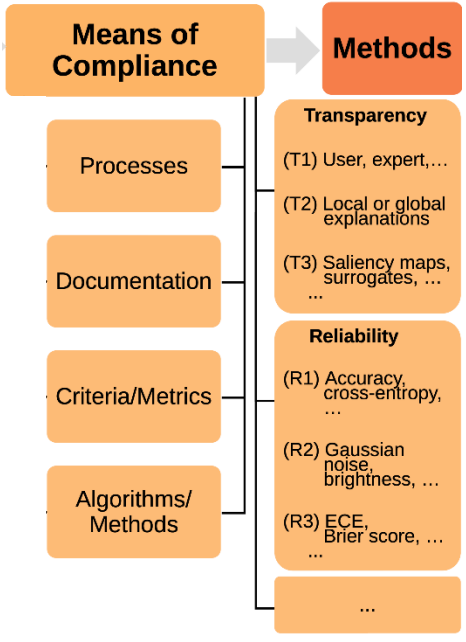
- Transparency objective (29 objectives)
  Example: Define transparency criteria for stakeholders
- Reliability objective (44 objectives):
  Example: Demonstrate robust behaviour against relevant perturbations



Certification Scheme: Means of Compliance and Methods

School of Engineering
IAMP Institute of Applied Mathematics and Physics

- Set of **means** to achieve **compliance with objectives** (29 + 44 objectives)
- Linked Means of Compliance (100 + 156 means of compliance)
- **Criteria/metrics for assessing means of compliance** (55 in total)
  - Qualitative criteria
  - Quantitative metrics
- **Algorithms/technical methods** (95 in total)
  Goal: providing compliance with objectives/criteria/metrics for different trustworthiness dimensions, e.g.
  - *Transparency*: LIME, SHAP, …
  - *Reliability*: perturbations, symbolic abstractions, ...



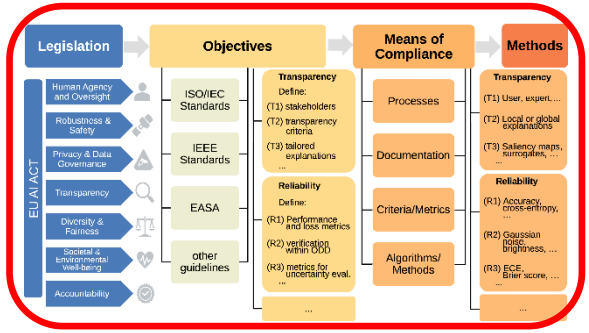
4. Example Use Case

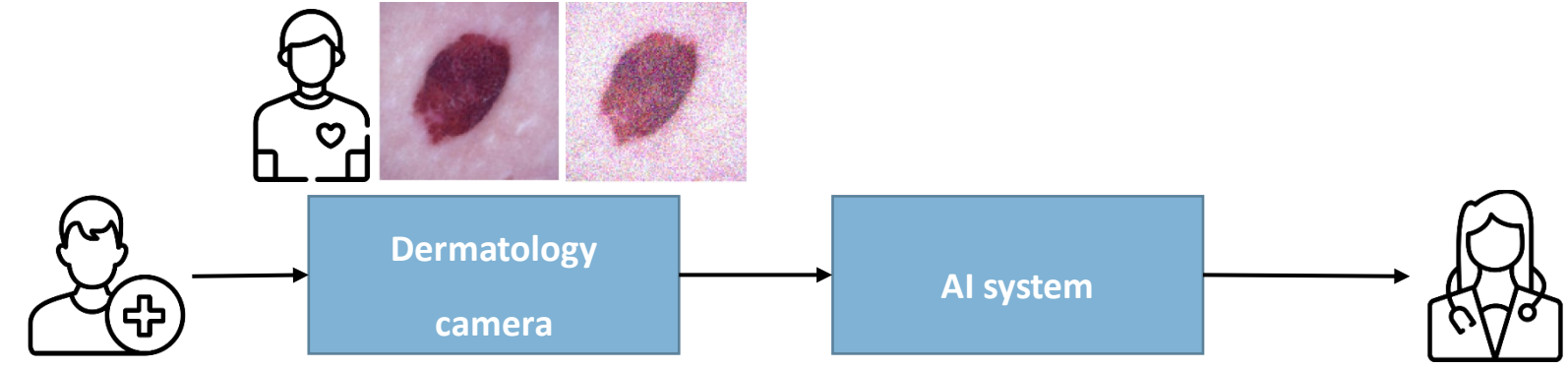
Use case

Third party use cases with industry (high and limited risk, computer vision) – **confidential**

Exemplary use case (high risk): **Skin lesion classification**

- Benign and malignant skin lesions

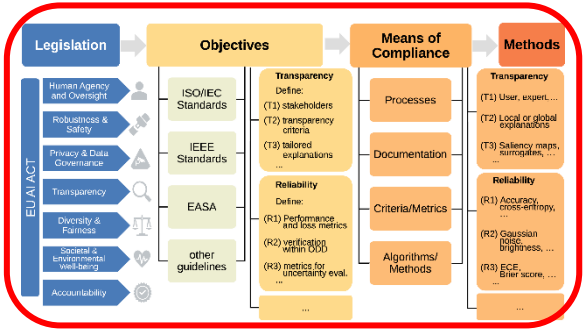


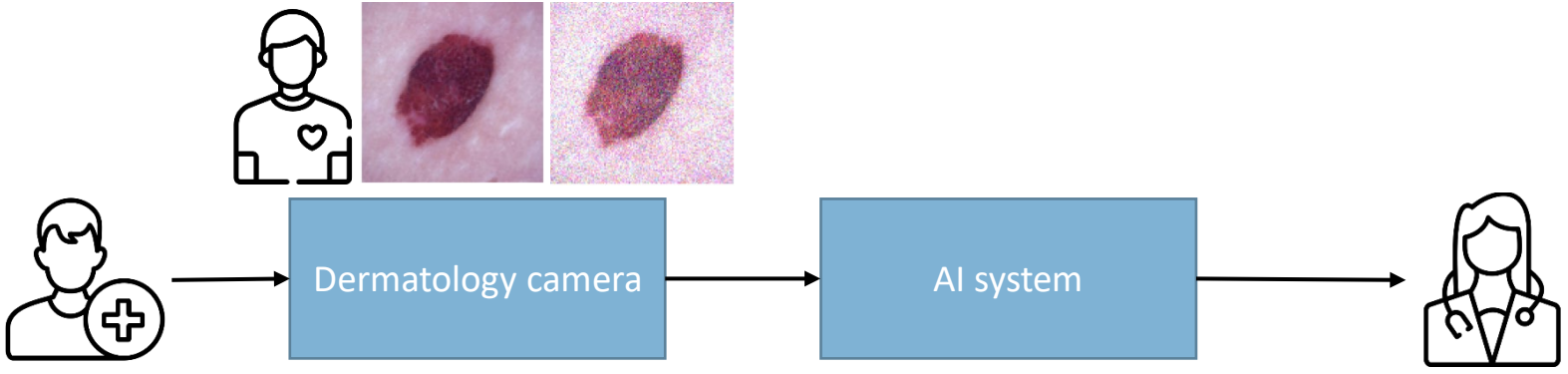


Use case

AI system must be **reliable** and **transparent**

- High risk
- **Need for certification** (compliance with objectives)





Reliability dimension

Reliability vs. Robustness

- **Reliability** = Property of **consistent intended** behaviour and results
- **Robustness** = Ability of an AI system to maintain its level of performance **under any circumstances**

Reliability includes **different aspects** within the certification scheme, e.g.:

- Data coverage
- Robustness
- Uncertainty

Reliability: Data coverage and Robustness

Objective (data coverage):
Data coverage of the Operational Design Domain (ODD)

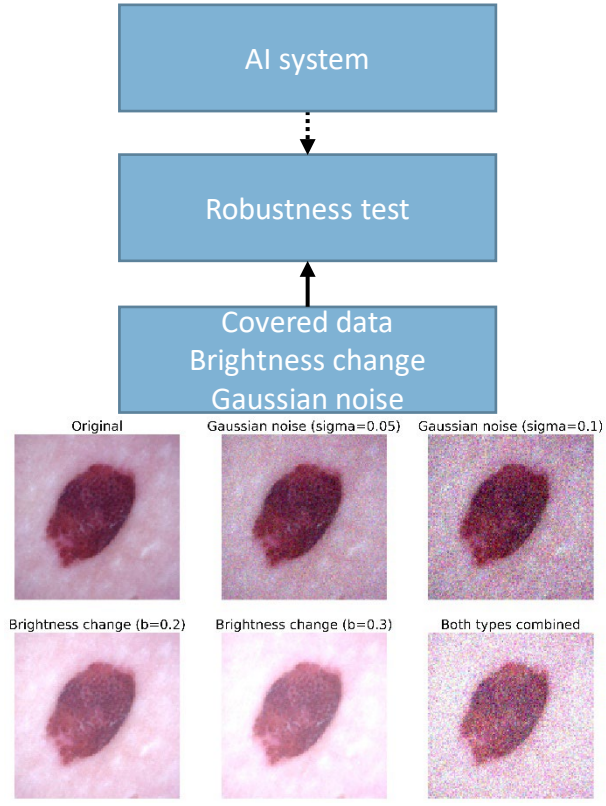
- ODD contains all application relevant perturbations including their intensity
- Area where the AI system must function reliably and robustly

Means of compliance:
Simulation of the relevant perturbations

Objective (robustness):
Robust behaviour against ODD relevant perturbations

Means of compliance:
Robustness tests

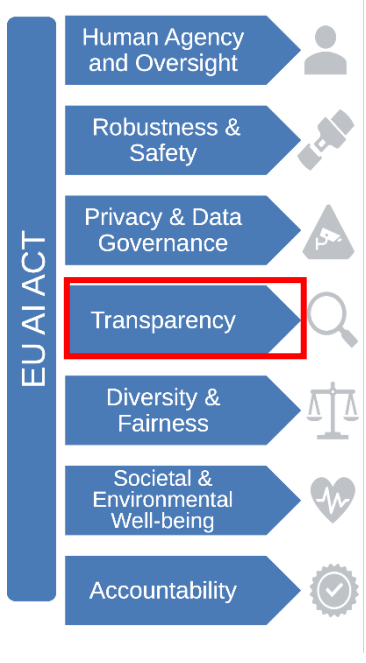
- Confidence interval (CI) tests
- Symbolic abstractions
- Exhaustive search (with constraints)



Robustness tests typically use the ODD data (original or simulated) as input to assess the robustness of the AI system

Transparency dimension

- Is the decision by the AI system **comprehensible**?
- What **stakeholders** are involved, what are their requirements?
- Does the AI system provide appropriate **explanations**?
  Pertaining to the data, model behaviour (**global**), or output (**local**)?
- Explanations: What **features** mainly caused a decision?

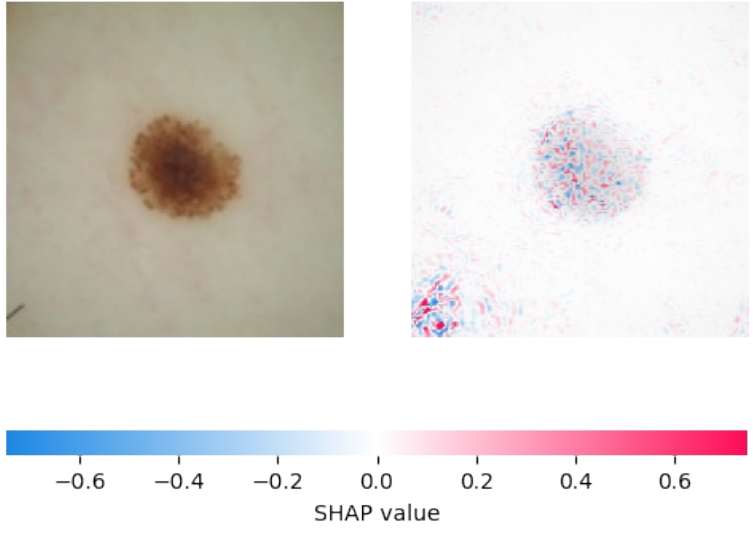


Transparency dimension: local explanations

Objective: “Provide the physician with explanation what lead to a malignant classification”

Means of compliance: Local explanations, e.g.

- saliency maps (gradient-based methods)
- SHAP

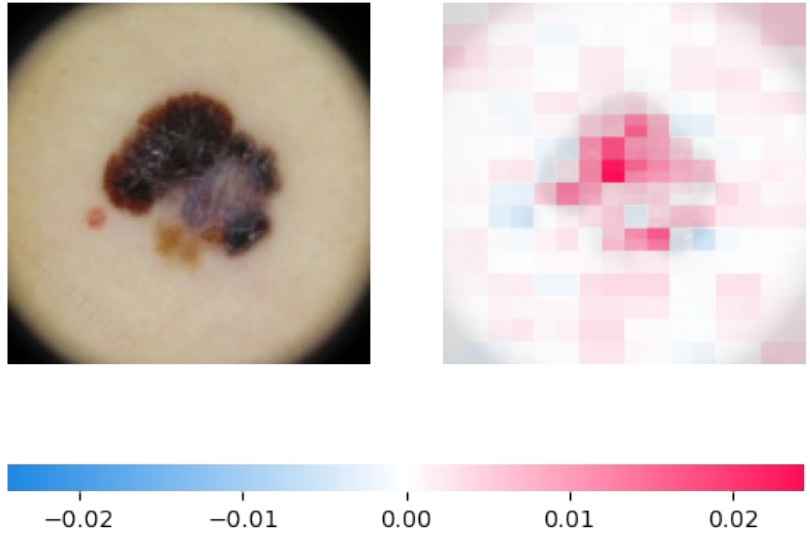


Explanation from SHAP using a gradient-based approach

Which explanation would a **physician** trust?

Malignant lesions typically have

- less defined border
- irregular shape and colour profile



Explanation from SHAP using a patch-based approach

Conclusion

School of Engineering
IAMP Institute of Applied Mathematics and Physics

- **Problem: gap** between high level regulations and technical assessment methods
- **Our Contribution:** a **methodology** with **actionable directives** for the certification of AI-based systems
  - so far: 73 objectives, 256 means of compliance, 55 metrics, 95 methods in certification scheme
- Methodology is generally applicable for **real-world** use cases;
- **Future Research Plans:**
  - **AI in safety-critical systems**: combination of new methods with classical methods from risk analysis and functional safety.
  - **Develop verification frameworks for AI in safety-critical systems** with research partners, tailored to the specific use case

Thank you!

About me:

Dr. Joanna Weng
Senior Lecturer ZHAW SoE
Safety Critical Systems Lab
Associate Faculty Member Centre for AI (CAI)
Contact: wenj@zhaw.ch

Innovation project supported by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Innosuisse – Swiss Innovation Agency

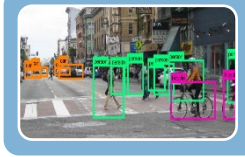
Centre for Artificial Intelligence (CAI)

School of Engineering
IAMP Institute of Applied Mathematics and Physics

Collaboration With CAI on AI-related projects

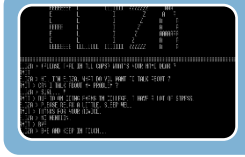
Autonomous Learning Systems

- *Reinforcement Learning*
- *Multi-Agent Systems*
- *Embodied AI*



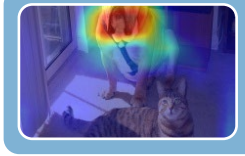
Computer Vision, Perception and Cognition

- *Pattern Recognition*
- *Machine Perception*
- *Neuromorphic Engineering*



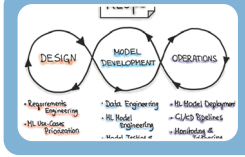
Natural Language Processing

- *Dialogue Systems*
- *Text Analytics*
- *Spoken Language Technologies*



Trustworthy AI

- *Explainable AI*
- *Robust Deep Learning*
- *AI & Society*



AI Engineering

- MLOps
- Data-Centric AI
- Continuous Learning

https://www.zhaw.ch/en/engineering/institutes-centres/cai/