Transparent, Safe, and Trustworthy AI

Enhancing Human-AI Collaboration in Critical Infrastructure

February 14th, 2025

AI4REALNET has received funding from European Union's Horizon Europe Research and Innovation programme under the Grant Agreement No 101119527

Introduction

Mohamed Hassouna – Fraunhofer IEE / University of Kassel

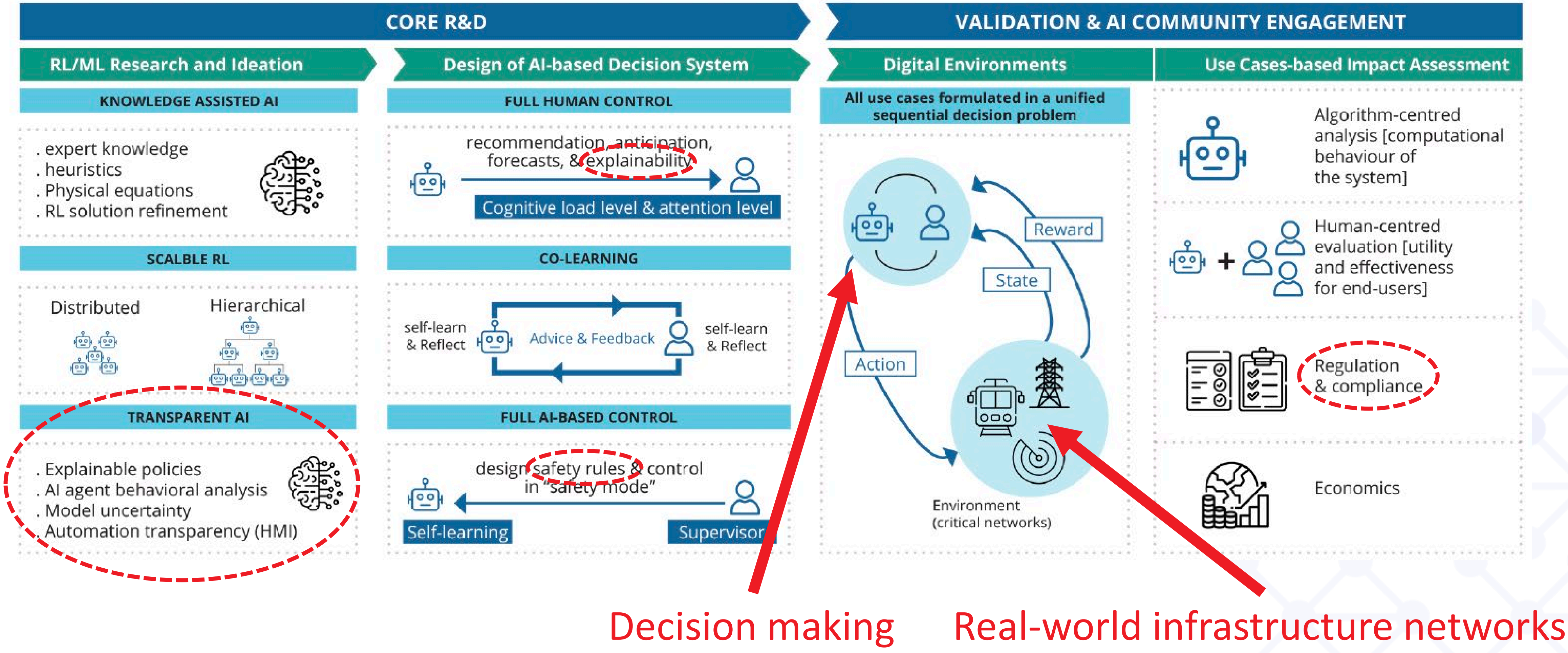
14.02.2025

AI4REALNET has received funding from European Union's Horizon Europe Research and Innovation programme under the Grant Agreement No 101119527

Welcome!

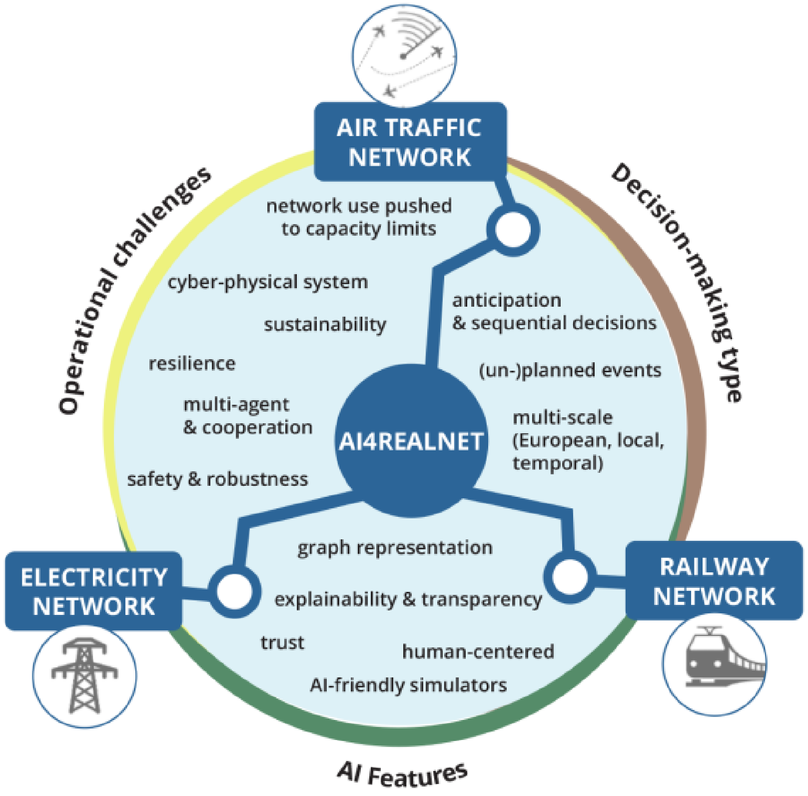
- Transparent, Safe, and Trustworthy AI - Enhancing Human-AI Collaboration in Critical Infrastructure
- Presenters:
  - Ricardo Chavarriaga – ZHAW
  - Alberto Maria Metelli – POLIMI
  - René Heinrich – Fraunhofer IEE
  - Clark Borst – TU Delft
  - Toni Wäfler – FHNW
- Organization:
  - Bianca Silva
  - Mohamed Hassouna

AI4REALNET Overview



AI4REALNET project objectives and scope

- Develop next generation of **decision-making methods** powered by supervised and reinforcement learning for **critical infrastructures**
- **Boost the development and validation of novel AI algorithms** via 3 existing open-source AI-friendly digital environments
- Three degrees of autonomy
  - AI-assisted human control
  - Human-AI co-learning
  - Autonomous AI



Project use cases: focus on critical infrastructures

UC1 POWER GRID — AI4REALNET

AI assistant supporting human operators' decision-making in managing power grid congestion

AI ROLE: Provide a human operator with remedial action recommendations aimed at safely managing overloads on the electrical lines and easing the workload of the human operator.

FULL HUMAN CONTROL — recommendation, anticipation, forecasts, & explainability — Cognitive load level & attention level

UC2 POWER GRID — AI4REALNET

Sim2Real, transfer AI-assistant from simulation to real-world operation

AI ROLE: Provide a human operator with remedial action recommendations, considering a transfer from training (digital) to real-world environments.

FULL HUMAN CONTROL — recommendation, anticipation, forecasts, & explainability — Cognitive load level & attention level

Rte — TenneT

DB — SBB CFF FFS

UC1 RAILWAY — AI4REALNET

Automated re-scheduling in railway operations

AI ROLE: The re-scheduling task is performed in a highly automated manner by an AI-based re-scheduling system. It observes the real-time state of all the trains and tracks in the control area of interest and automatically detects the need to intervene, decides on an intervention, and executes this intervention.

FULL AI-BASED CONTROL — design safety rules & control in "safety mode" — Self-learning — Supervisor

UC2 RAILWAY — AI4REALNET

AI-assisted human re-scheduling in railway operations

AI ROLE: Assist the human dispatcher in railway operations in re-scheduling train runs to fulfil all offered services and minimize delays for the customer.

JOINT DECISION MAKING — self-learn & Reflect — Advice & Feedback — self-learn & Reflect

NAV PORTUGAL — TU Delft

UC1 ATM — AI4REALNET

Airspace sectorization assistant

AI ROLE: Partially and fully automate the sectorization process to assist or replace the staff manager in deciding when and how to split and merge sectors to balance the workload of tactical ATCOs.

FULL HUMAN CONTROL — recommendation, anticipation, forecasts, & explainability — Cognitive load level & attention level

UC2 ATM — AI4REALNET

Flow and airspace management assistant

AI ROLE: Provide advice to air traffic controller about deviations with better sector capacity adherence and performance measured by an indicator of environmental area. Also consider the need to review the sectorization plan due to the activation of military areas and required trajectory efficient deviations.

JOINT DECISION MAKING — self-learn & Reflect — Advice & Feedback — self-learn & Reflect

AI4REALNET project objectives and scope

- **Central role of human → Trustworthiness**
- **Regulatory and safety requirements**
  - Regulations and safety standards demand systems to align with predefined rules and guidelines
- **Need for Explainability**
  - Operators and stakeholders must trust AI decisions, which necessitates explainability
  - trace and justify AI actions
- **Transparency**
  - Design of AI decision systems with transparency in mind
  - Enhance understanding & usability, reduce cognitive stress, ..

Outline

- Introduction (Mohamed Hassouna – Fraunhofer IEE / University of Kassel)
- Assessing trustworthiness and regulatory compliance (Ricardo Chavarriaga – ZHAW)
- Safe Reinforcement Learning (Alberto Maria Metelli – POLIMI)
- Explainable AI (René Heinrich – Fraunhofer IEE)
- Designing for Transparency (Clark Borst – TU Delft)
- Human Agency (Toni Wäfler – FHNW)
- Q&A

Assessing trustworthiness and regulatory compliance

Ricardo Chavarriaga

EU Artificial Intelligence Act

EU AI ACT

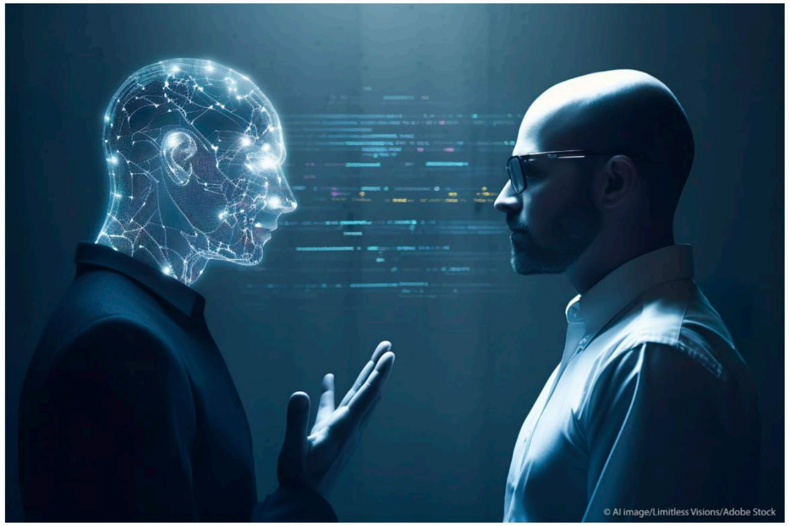
- Human Agency and Oversight
- Technical Robustness & Safety
- Privacy & Data Governance
- Transparency
- Diversity, Non-Discrimination & Fairness
- Societal & Environmental Well-being
- Accountability

European Parliament

EU AI Act: first regulation on artificial intelligence

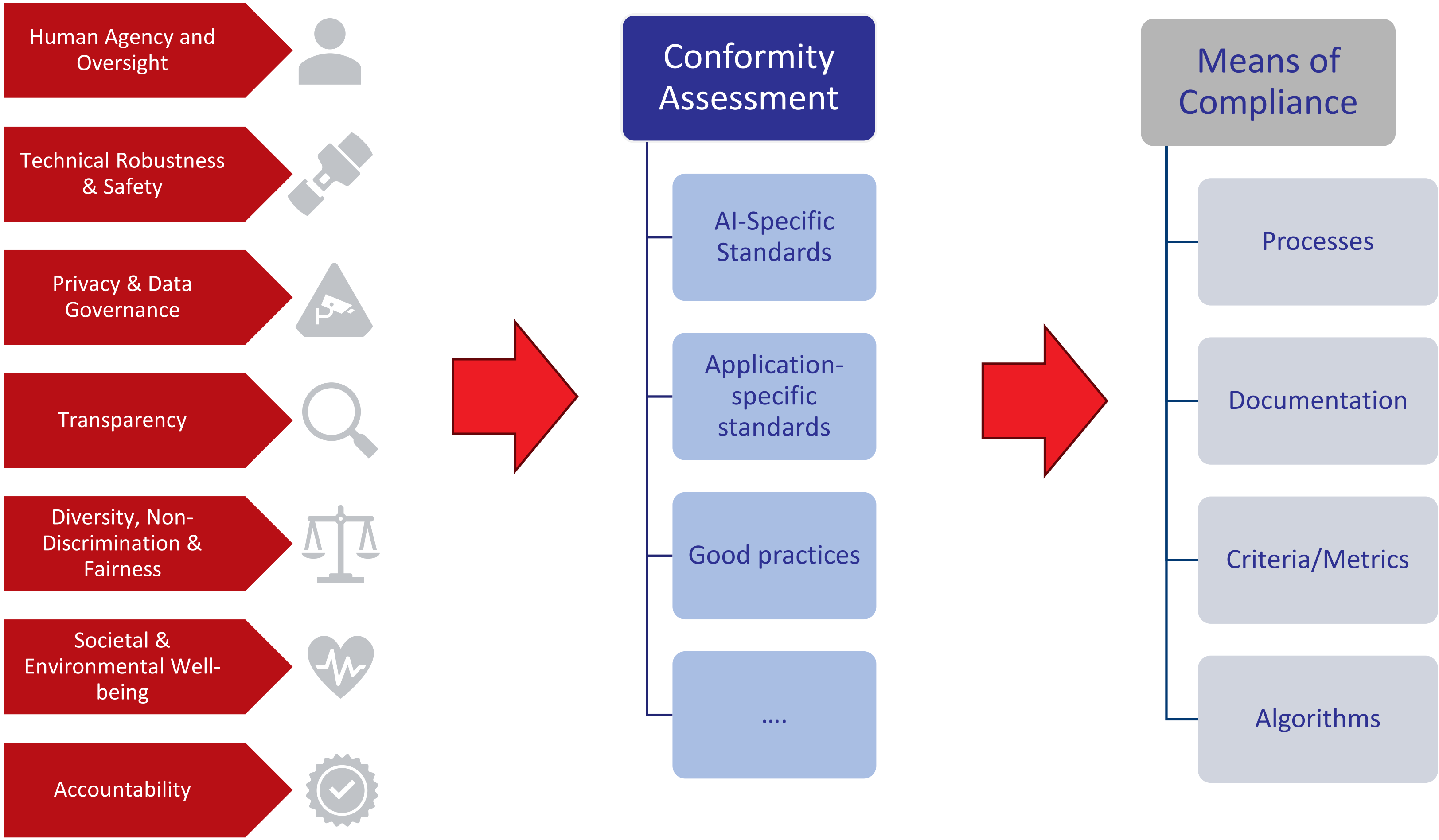
Society Updated: 14-06-2023 - 14:06
Created: 08-06-2023 - 11:40

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.



This illustration of artificial intelligence has in fact been generated by AI

Assessing trustworthiness/regulatory compliance



However…

Punctual Ex-post assessment is not enough….

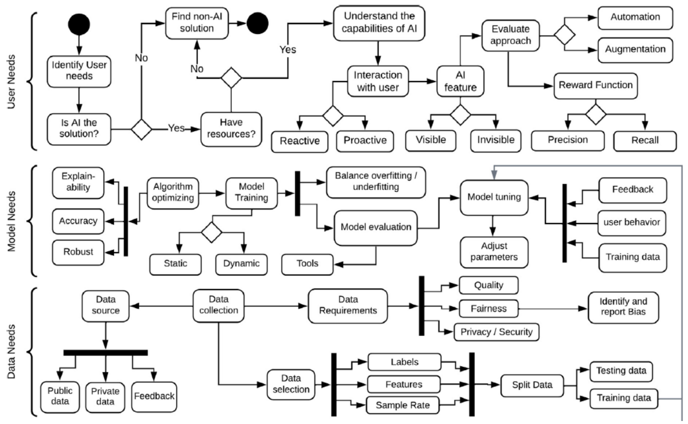
Impact and risk assessment to be implemented throughout the entire life-cycle:

- “risk management system shall be established, implemented, documented and maintained”. (EU AI Act Art 9.)

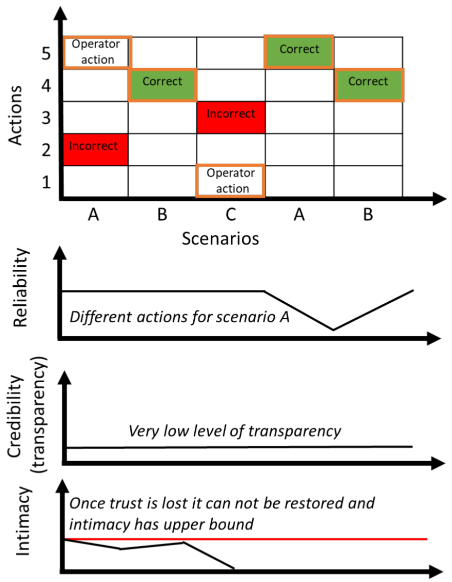
Key elements

- Identifying functional and non-functional requirements and KPIs related to trustworthy dimensions
- Adopting risk management approaches that are suitable for AI-related risks and safety critical systems

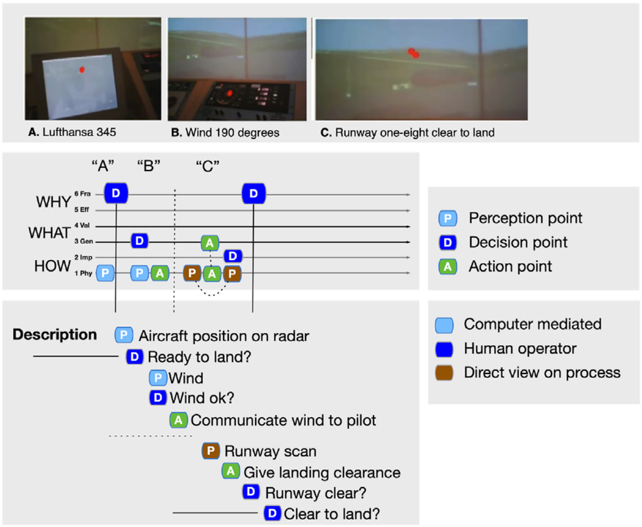
Alternatives – Descriptive Frameworks



[Requirements Engineering for Human-centered AI]



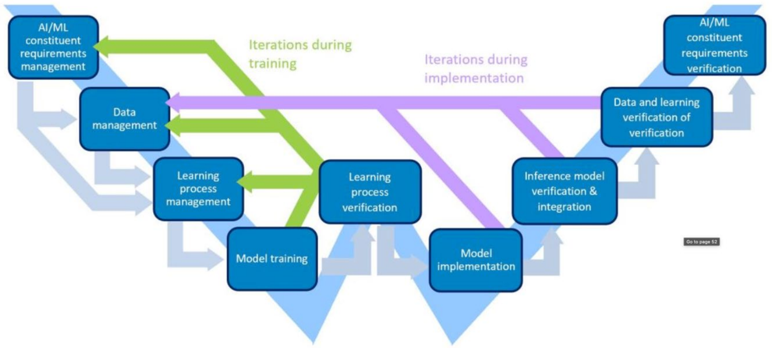
[L2RPN Framework]



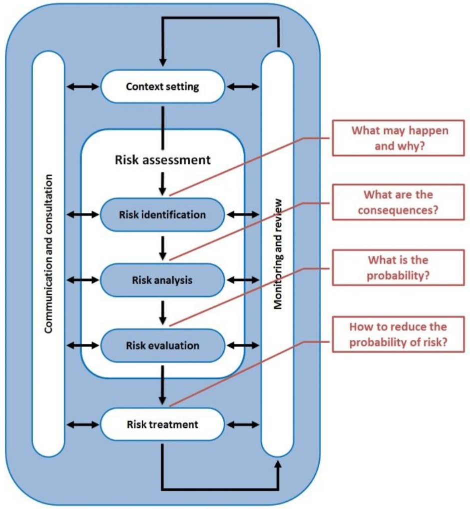
[Joint Control Framework]

Alternatives – Risk Management Frameworks

NIST RMF

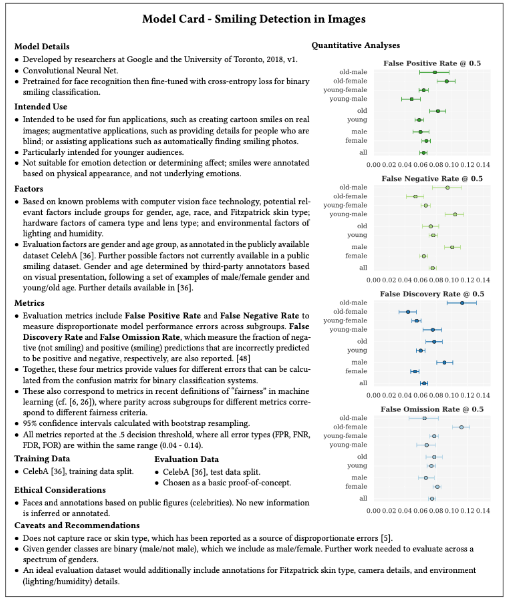


EASA Guidance for ML applications



ISO/IEC Standards 31000/ 23984

Alternatives – Qualitative/Quantitative Assessments



Model Card - Smiling Detection in Images

Model Details
- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use
- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors
- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics
- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

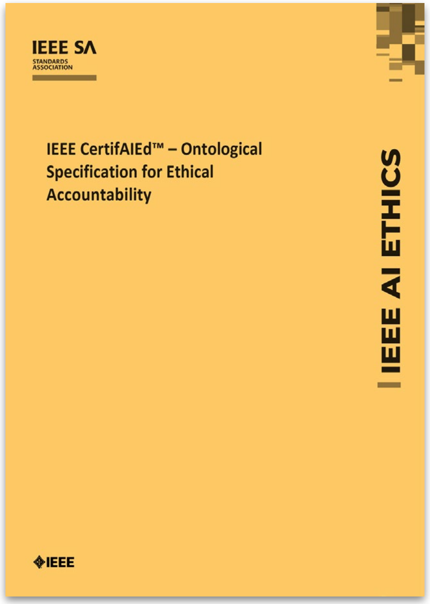
Training Data
- CelebA [36], training data split.

Evaluation Data
- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses



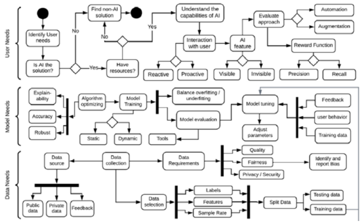
Model Cards
Datasheets for Datasets

IEEE CertifAIEd

EU Assessment List
Trustworthy AI
ALTAI

Alternatives

Descriptive Frameworks



Not always focused on ethical/social aspects.
Too technical. Hard to define a project-wide framework at this stage

Risk Management Frameworks

Not aligned with regulatory demands (bar ISO/IEC)
Based on classical risk management approaches
Partially suited for AI-related risks

Assessment instruments

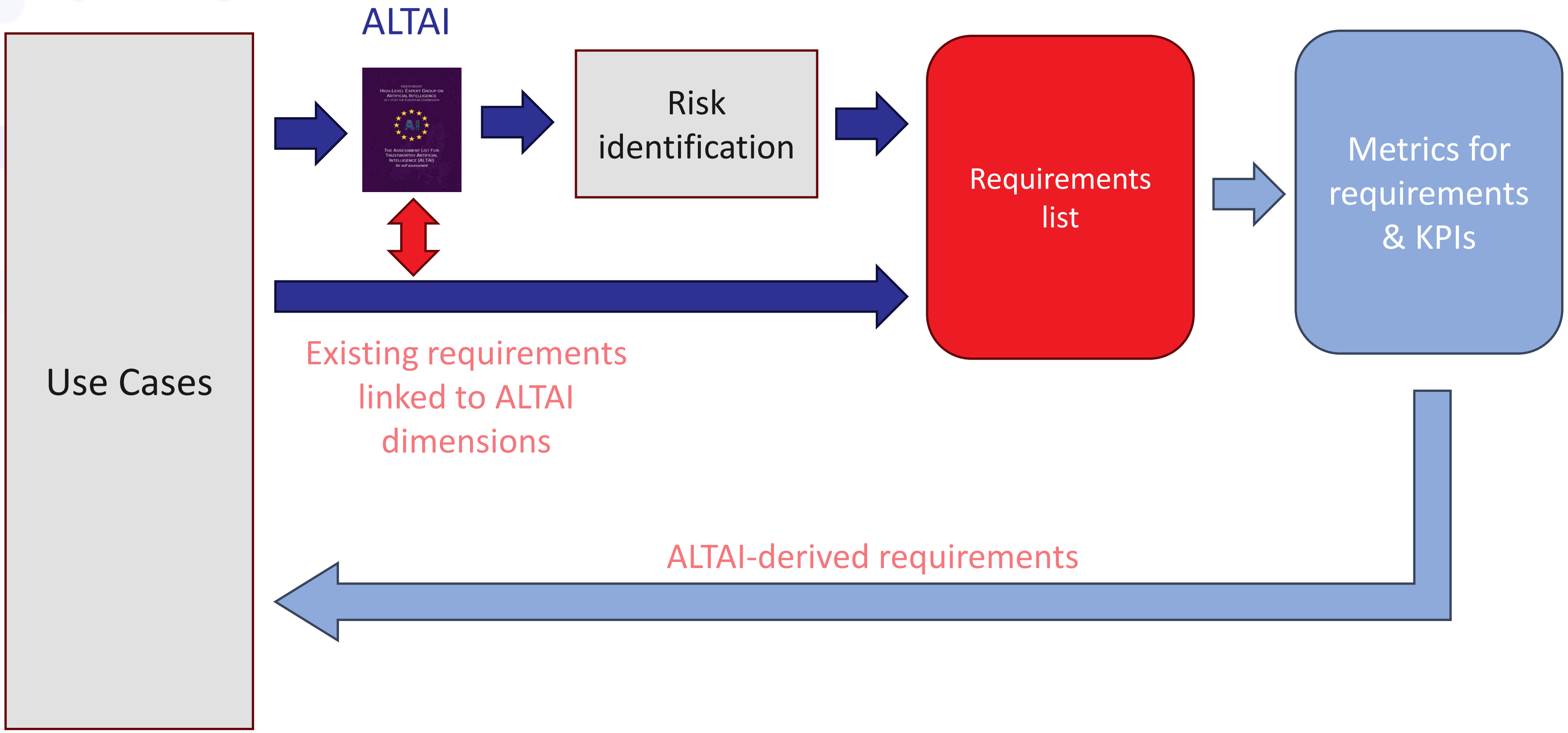
Focused on ethical/societal aspects
Designed for <u>ex-post</u> analysis

ALTAI has a common basis with AI act

ALTAI – Assessment List for Trustworthy AI

- **Human Agency and Oversight**
  - Human Agency & Autonomy
  - Human Oversight
- **Technical Robustness & Safety**
  - Resilience to Attacks & Security
  - General Safety
  - Accuracy
  - Reliability, Fall-back plans & Reproducibility
- **Privacy & Data Governance**
  - Privacy
  - Data Governance
- **Transparency**
  - Traceability
  - Explainability
  - Communication
- **Diversity, Non-Discrimination & Fairness**
  - Avoidance of Unfair Bias
  - Accessibility & Universal Design
  - Stakeholder Participation
- **Societal & Environmental Well-being**
  - Environmental Well-being
  - Impact on Work and Skills
  - Impact on Society at large & Democracy
- **Accountability**
  - Auditability
  - Risk Management

Process



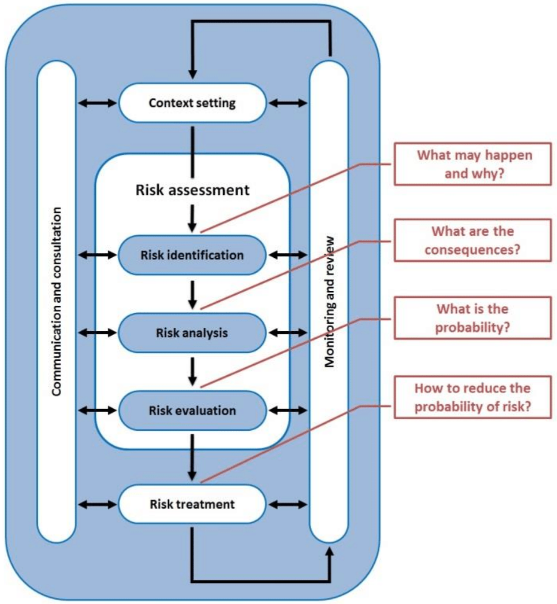
The outcome of this process is reported in the AI4REALNET Use Case definitions, available in the project website.

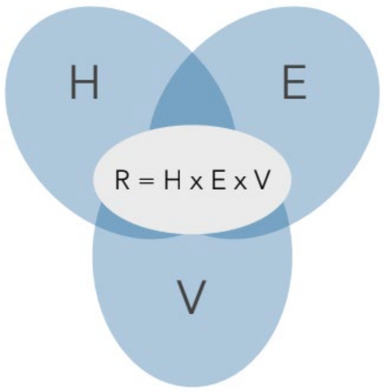
Work in progress

- Design of ALTAI-compatible approaches for continuous assessment from ex-post to continuous assessment
- Specific to safety critical systems
  - Adapt the questionnaire for ex-ante and continuous analysis
  - Link to descriptive and system-level frameworks

Work in progress

- Integration with risks management frameworks
  - Define metrics for ethically-relevant dimensions
  - Ethically-informed **multi-component approach to risk analysis (Hazard, Exposure, Vulnerability)**



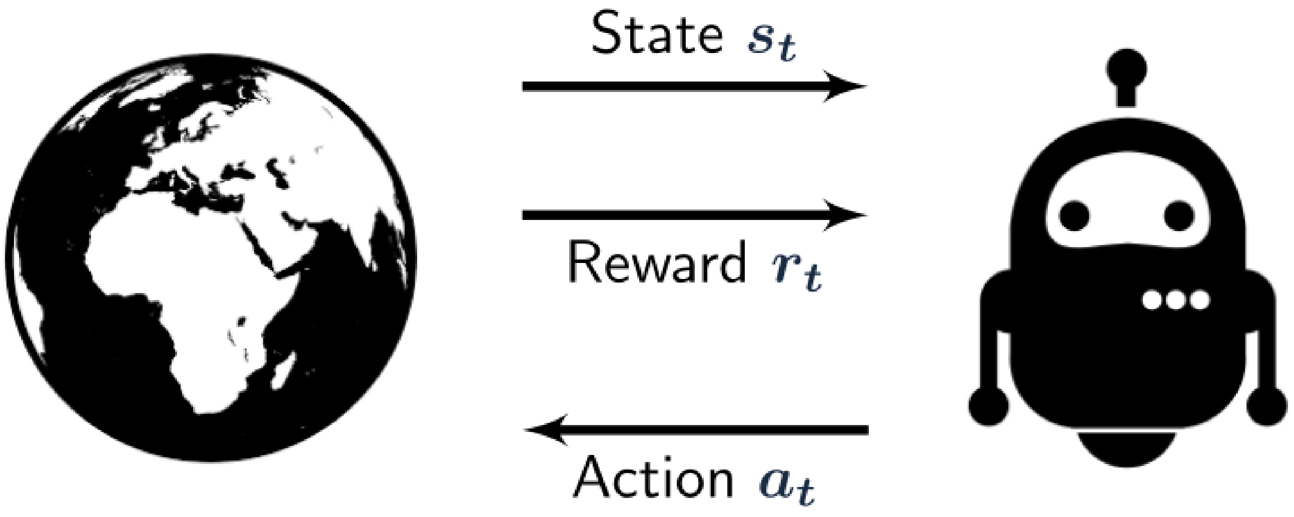


Safe Reinforcement Learning

Alberto Maria Metelli

Reinforcement Learning (RL)

- Sequential decision-making under uncertainty
- **Goal**: learn a **policy** maximizing the **expected return**, i.e., expected cumulative sum of the rewards



$$\max_{\pi \in \Pi} E_\pi\left(\sum_{t=0}^{\infty} \gamma^t r_t\right),$$

Sutton, Richard S., and Andrew G. Barto. "Reinforcement learning: An introduction." *2018.*

Different Definitions of Safe RL

Two facets of Safe RL:

- Safety **at the end** of the learning process
  - The safety caharacteristics have to be ensured by the policy produced at the end of the learning process
  - No interest in the safety of the policies played during the learning process

- Safety **during** the learning process
  - The safety characteristics have to be ensured by all the policies encountered during the learning process
  - This limits the exploration

Garcıa, Javier, and Fernando Fernández. "A comprehensive survey on safe reinforcement learning." *Journal of Machine Learning Research* 16, no. 1 (2015): 1437-1480.

Safety at the end of the learning process

- Standard RL optimizes the expected return
- Safety requires modifying the **optimization criterion**:
  - **Robust RL**: we have uncertainty on the environment parameters
  - **Risk-sensitive RL**: we have to take into account the stochasticity of the process in the learning objective
  - **Constrained RL**: we have to satisfy constraints defined in terms of costs/rewards

García, Javier, and Fernando Fernández. "A comprehensive survey on safe reinforcement learning." *Journal of Machine Learning Research* 16, no. 1 (2015): 1437-1480.

Robust RL

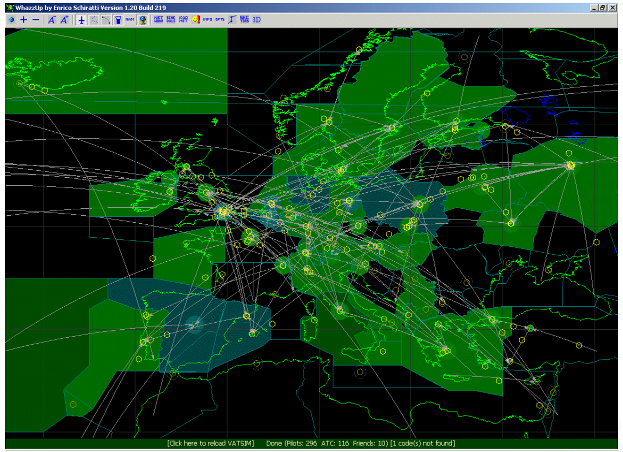
- Used to model uncertainty in the environment parameter
  - E.g., we don't know which environment we will face (**uncertainty set**)
- **Idea**: maximize the **worst-case expected return**, i.e., the expected return in the most challenging environment
- Can be formulated as min-max game
- Efficient solution for specific models of uncertainty (rectangular)

$$\max_{\pi \in \Pi} \min_{w \in \Omega^{\pi}} E_{\pi,w}\left(\sum_{t=0}^{\infty} \gamma^t r_t\right),$$

Morimoto, Jun, and Kenji Doya. "Robust reinforcement learning." *Neural computation* 17, no. 2 (2005): 335-359.

Risk-sensitive RL

- Used when we want to be safe w.r.t. the stochasticity of the environment
  - E.g., we want to guarantee a minimum gain with a certain probability (financial scenarios)
- **Idea**: maximize the **risk-sensitive** transformation of the **expected return**
  - Mean-variance
  - CVaR
  - Volatility



$$\max_{\pi \in \Pi} E_\pi(R) + \frac{\beta}{2} Var(R)$$

Mihatsch, Oliver, and Ralph Neuneier. "Risk-sensitive reinforcement learning." *Machine learning* 49 (2002): 267-290.

Constrained RL

- Used when we have constrained to be satisfied
  - E.g., limit the magnitude of the action, control how many times a certain region is visited
- **Idea**: maximize the **expected return subject to constraints**
  - Modeled with the formalism of the **Constrained Markov Decision Processes**
- Can be addressed using the Lagrangian approach to turn it into a min-max unconstrained optimization

$$\max_{\pi \in \Pi} \; E_\pi(R) \; subject \; to \; c_i \in C, c_i = \{h_i \leq \alpha_i\},$$

Altman, Eitan. *Constrained Markov decision processes*. Routledge, 2021.

Safety during the learning process

The safety characteristic has to be ensured **at every step of the learning process**

- **Safe exploration**: the performance must always remain above a fraction of the performance of a baseline one
    - The more we want to be close to the baseline, the less we explore
    - We may not reach the global optimum
- **Monotonic performance improvement**: the performance must always non-decrease
    - Makes the learning process slower, but smoother

Papini, Matteo. "Safe policy optimization." (2020).

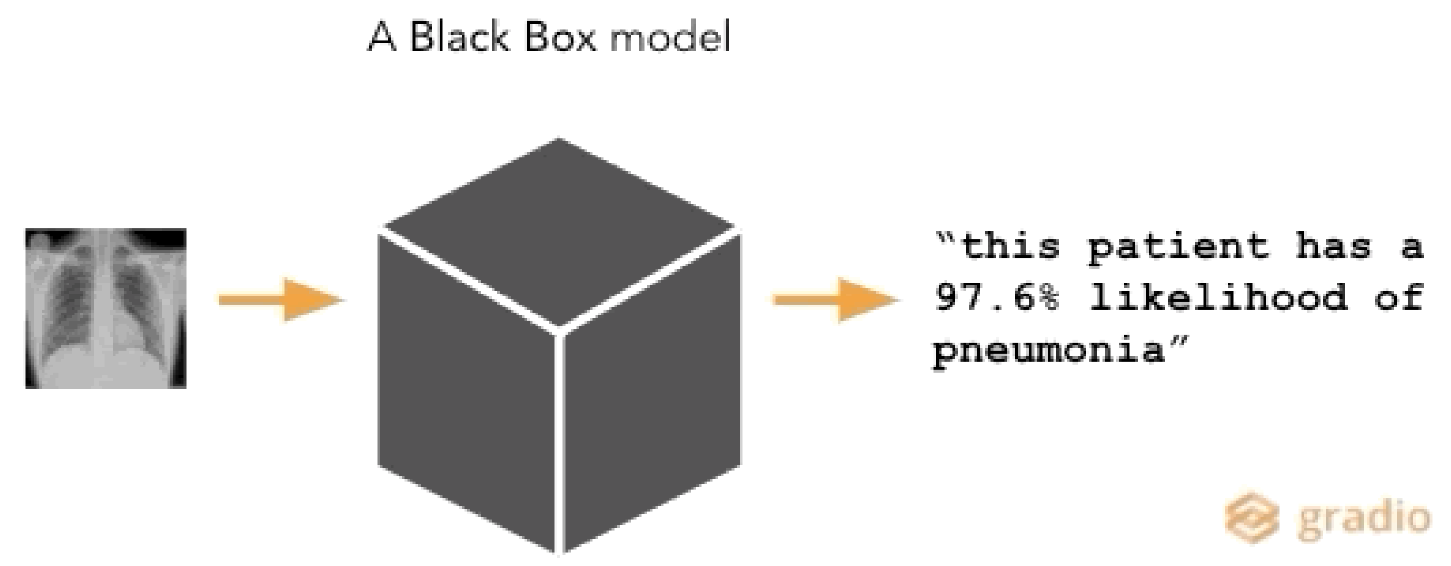
Explainable AI

René Heinrich

What is Explainable Artificial Intelligence?

Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.

Source: Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information fusion 58 (2020): 82-115.



Source: https://towardsdatascience.com/bridging-the-interpretability-gap-in-medical-machine-learning-66bdf1446a4a

Categorization of Explanation Methods

1. Type of Explanation

Post-hoc Explainability

- Problem:
  How can black-box models be explained?

→ Application of methods to analyze the model after training.

Inherently Interpretable Models

- Problem:
  How can transparent models be designed?

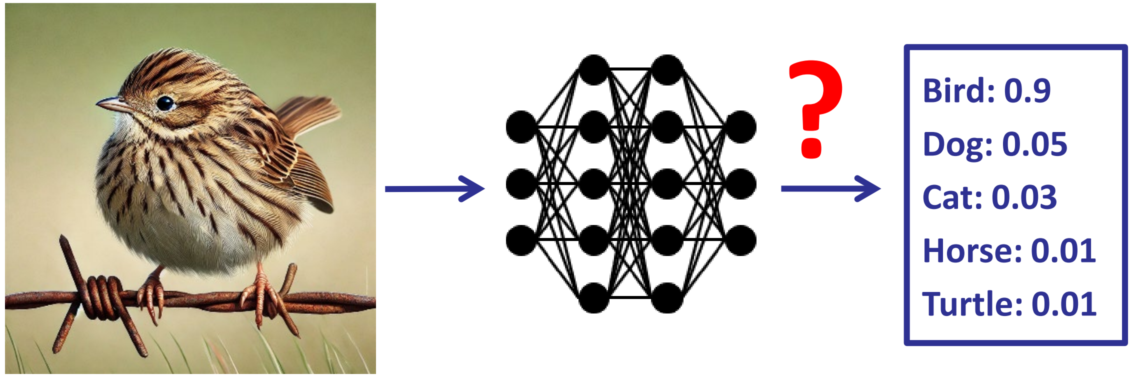
→ Limiting model complexity.

Categorization of Explanation Methods

2. Scope of Explanation

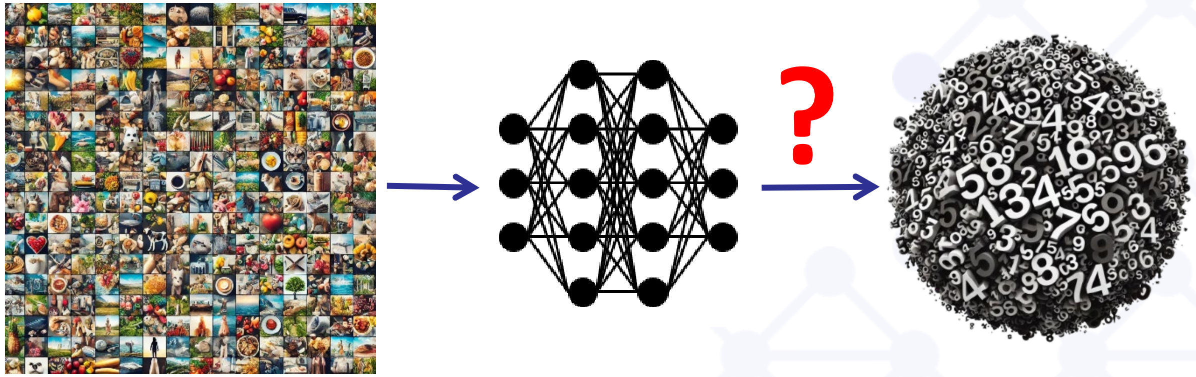
Local Explainability

- Explanation of <u>individual predictions</u> of a model.
    - What features are particularly important?
    - How do features influence the prediction?
    - How must features be changed to predict a different value?



Global Explainability

- Explanation of the <u>entire model behavior</u>.
    - What features are particularly important?
    - How does the model make its decisions?
    - What concepts has the model learned?



Categorization of Explanation Methods

3. Format of Explanation

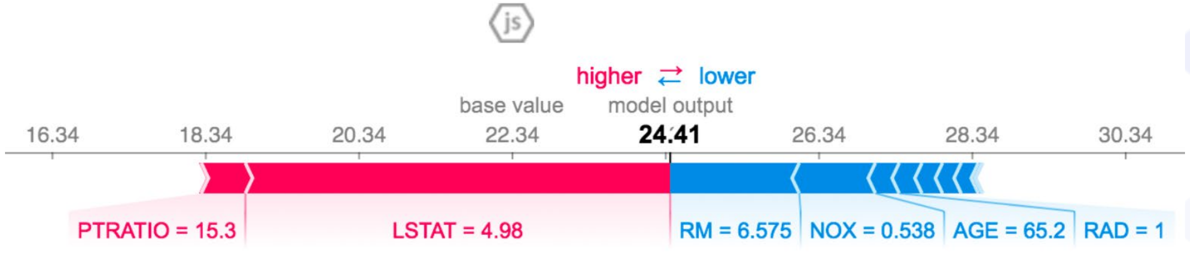
1) Summary Statistics

- e.g., a numerical value for feature importance

Feature A: 3.0, Feature B: 1.5, Feature C: 0.1

2) Visualizations

- e.g., impact of different features



Source: https://raw.githubusercontent.com/slundberg/shap/master/docs/artwork/boston_instance.png

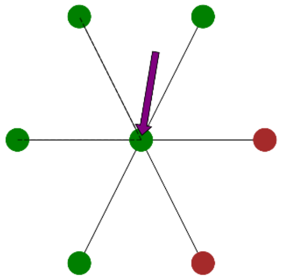
3) Text

- Generation of explanations in textual form

This image was classified as a "zebra" due to its black-and-white stripe pattern.

4) Example Data Points

- Extraction or generation of representative examples



Different Target Groups for Explanations

Each target group requires different explanations!

Target groups of Explainable AI

Data scientists, developers, product managers
- Ensuring and improving model quality
- Faster debugging
- New functionalities

Regulatory authorities
- Proof of legal compliance
- Auditing

Domain experts / users of the model
- Trust
- Scientific insights

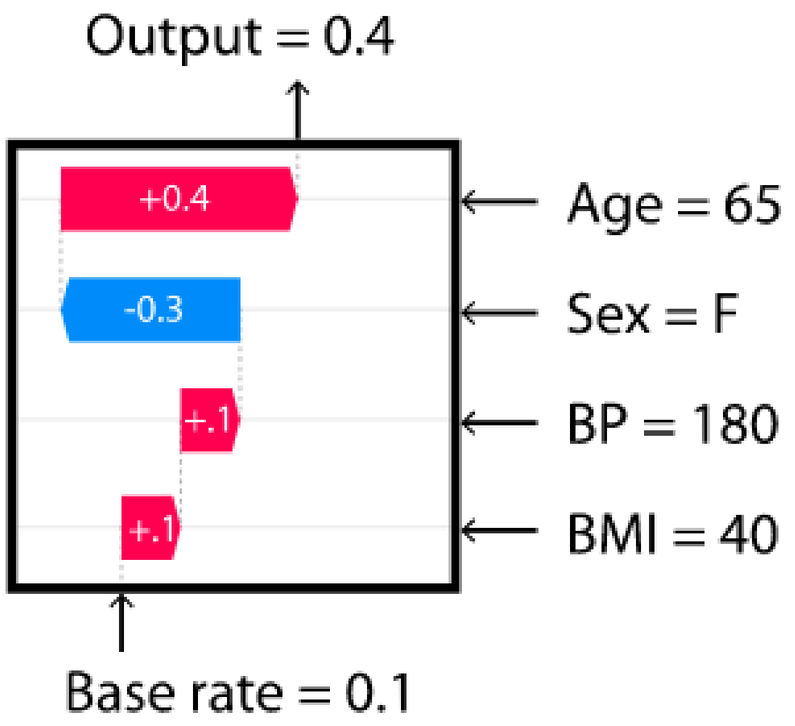
Users affected by model decisions
- Understanding of their situation
- Fairness of decisions

Managers and executives
- Assessment of compliance
- Understanding of AI applications within the company

Source: Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information fusion 58 (2020): 82-115.

SHAP - SHapley Additive exPlanations

- Methods to approximate Shapley values
- Calculates how each feature contributes to a prediction
  → Local explanations
- Aggregation of local explanations enables global explanations
- Unifies many other post-hoc explanation methods
  (e.g., LIME, LRP, DeepLIFT)



Source: https://github.com/slundberg/shap

SHAP - SHapley Additive exPlanations

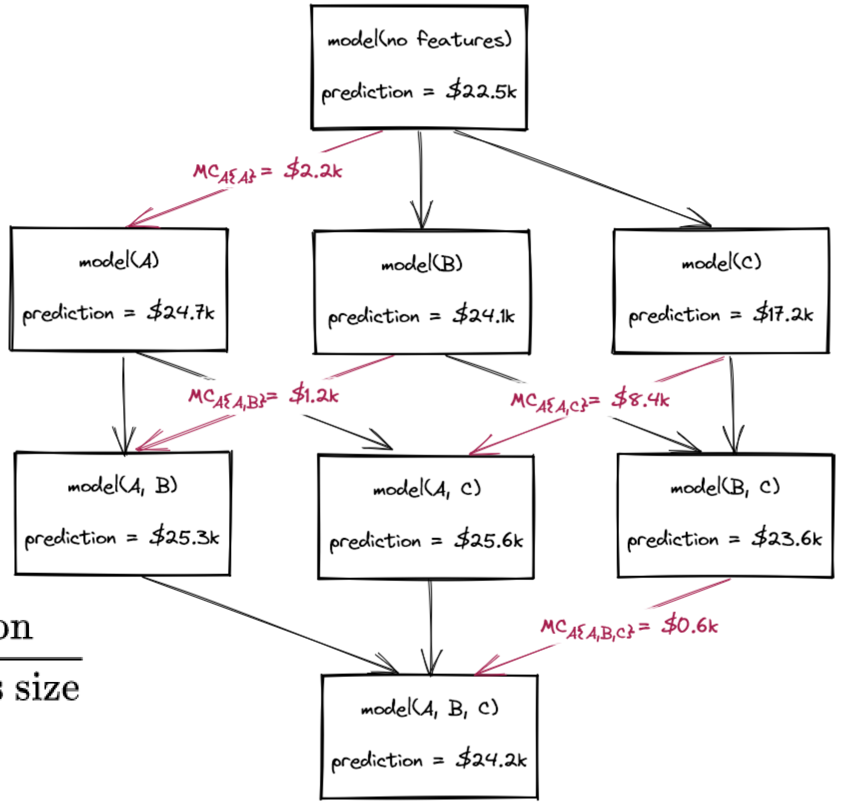
Shapley Values

- Game-theoretic method for distributing the gains of a cooperative game among a group of players
- Can also be used for explaining ML models
- Calculation of the Shapley value for a feature i:

$$\varphi_i(v) = \frac{1}{n} \sum_{S \subseteq N \setminus \{i\}} \binom{n-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S))$$

$$\varphi_i(v) = \frac{1}{\text{number of players}} \sum_{\text{coalitions excluding } i} \frac{\text{marginal contribution of } i \text{ to coalition}}{\text{number of coalitions excluding } i \text{ of this size}}$$

Source: https://en.wikipedia.org/wiki/Shapley_value



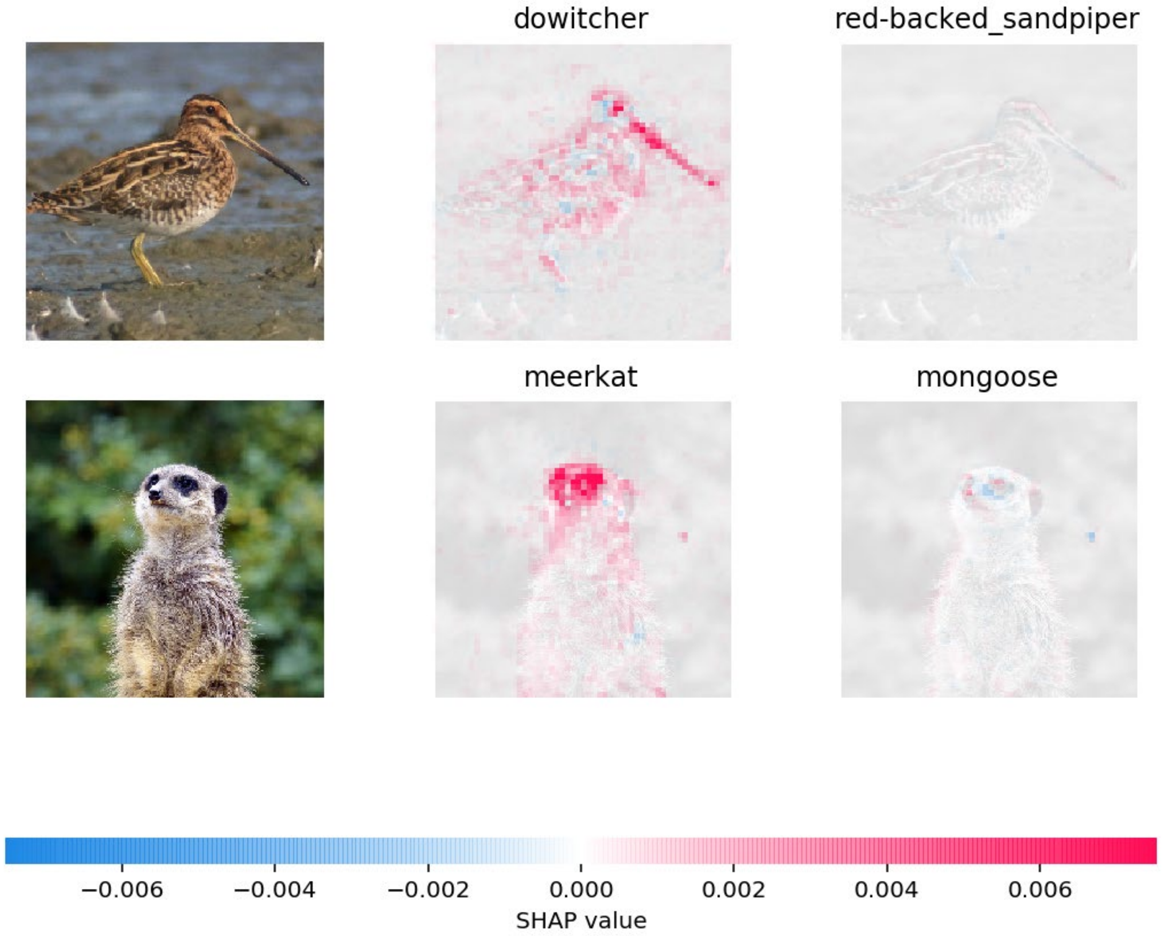
Source: https://www.aidancooper.co.uk/how-shapley-values-work/

- **Problem:** Calculation is NP-hard (especially with many features, computation times can be very long)

SHAP - SHapley Additive exPlanations

Advantages of SHAP:

- Model-agnostic
- Suitable for various data types (tabular data, image data, etc.)
- Solid mathematical foundation
- Fast approximation of Shapley values



Source: https://github.com/slundberg/shap

Disadvantages of Post-Hoc Methods

PERSPECTIVE

https://doi.org/10.1038/s42256-019-0048-x

nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Source: Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature machine intelligence 1.5 (2019): 206-215.

Disadvantages of Post-Hoc Methods

- **Approximation errors**: Post-hoc methods provide explanations that do not fully match the model's calculations.
- **Reduced trust**: An inaccurate explanation model reduces trust in both the explanation and the black-box model it tries to explain.
- **Misleading explanations**: Often misleading or insufficiently detailed to understand what a black-box model does.
- **Black-box models are usually unnecessary**: Their accuracy is typically not better than well-designed interpretable models

  → The belief in a trade-off between accuracy and interpretability is a misconception!

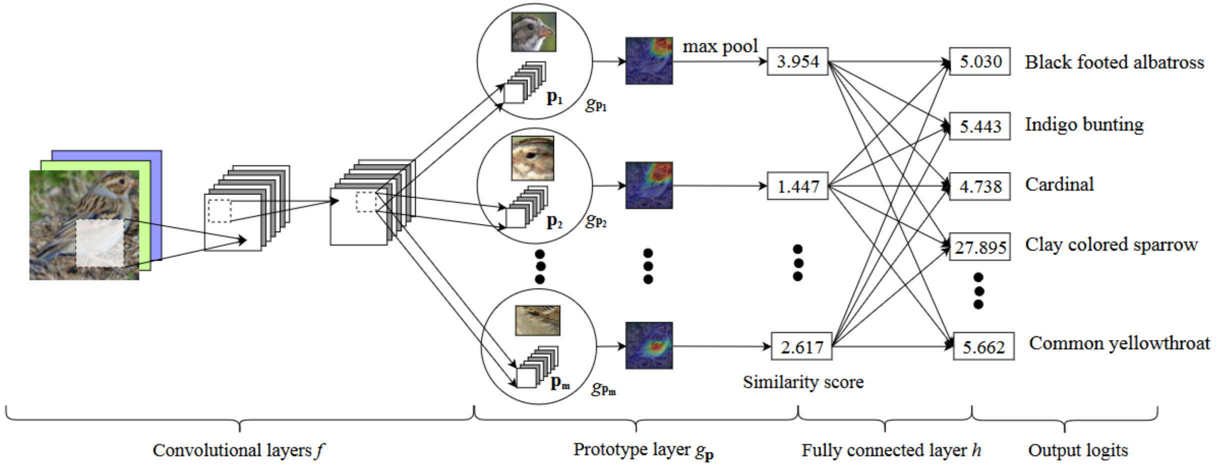
Interpretable Artificial Intelligence

- Model architectures inherently designed to be interpretable.
- Examples: Linear Regression, Logistic Regression, Decision Trees, KNN, etc.
- Often performs just as well as black-box models when dealing with structured data and meaningful features.
- **Problem:** Deep learning usually performs better for unstructured data (audio, image, etc.).

→ Development of interpretable deep learning models required!

Deep Prototype Learning

- **Concept:** Uses deep learning for feature extraction, but predictions are based on an interpretable combination of extracted features.
- **Training:** Identifies representative training image parts as prototypes for each class.
- Prediction for a new test image
  1. Find parts of the test image similar to learned prototypes.
  2. Classify based on weighted similarity to prototypes.

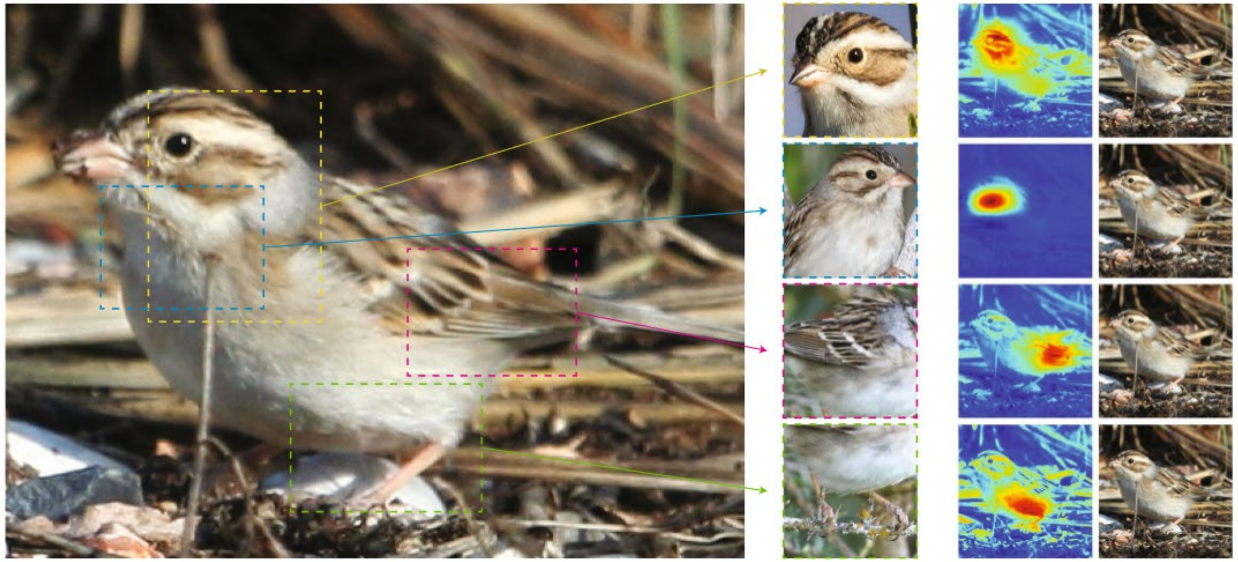


Source: Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." Advances in neural information processing systems 32 (2019).

Deep Prototype Learning

- Provides explanations in the form:
  “This bird is a yellowhammer because its head resembles the prototypical head of a yellowhammer, and its wings resemble the prototypical wings of a yellowhammer.”

→ Explanation of decisions in a manner similar to how experts classify images.



Source: Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature machine intelligence 1.5 (2019): 206-215.

Deep Prototype Learning

- The deep prototype learning model is capable of learning prototypical parts of 200 bird classes.
- The model's classifications are similarly accurate as those of non-interpretable black-box models.

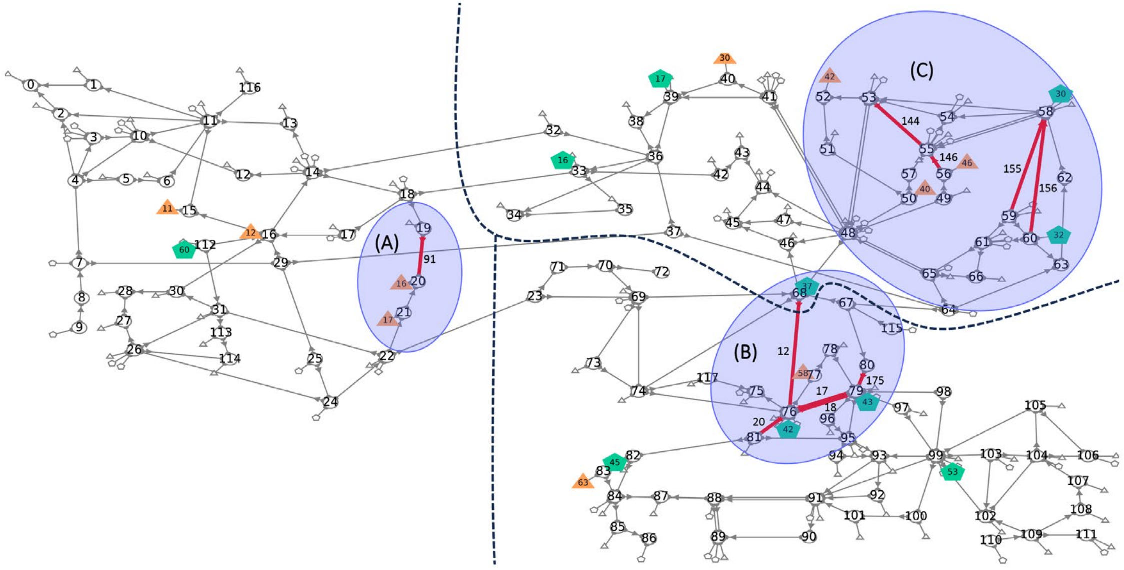
Table 1: Top: Accuracy comparison on cropped bird images of CUB-200-2011
Bottom: Comparison of our model with other deep models

Base	ProtoPNet	Baseline	Base	ProtoPNet	Baseline
VGG16	76.1 ± 0.2	74.6 ± 0.2	VGG19	78.0 ± 0.2	75.1 ± 0.4
Res34	79.2 ± 0.1	82.3 ± 0.3	Res152	78.0 ± 0.3	81.5 ± 0.4
Dense121	80.2 ± 0.2	80.5 ± 0.1	Dense161	80.1 ± 0.3	82.2 ± 0.2

Source: Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." Advances in neural information processing systems 32 (2019).

Application to Power Grid Usecases

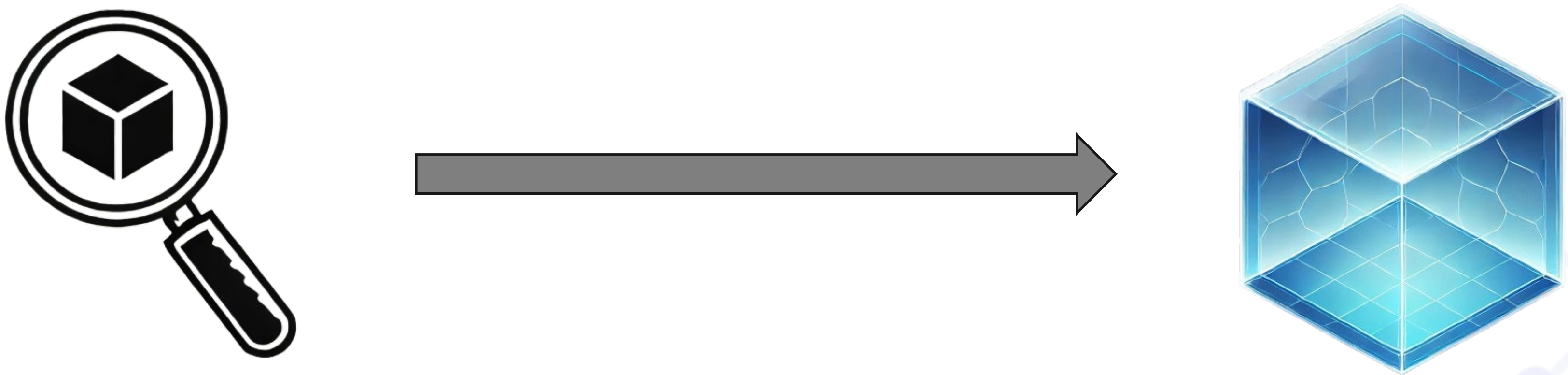
- **Visualization**: Feature importance is mapped onto the grid, allowing operators to identify influential elements.
- The power grid naturally forms a graph structure, applying and adjusting XAI methods to Graph-based agents (GNN) is needed.
- Adjust Prototype Learning to learn prototypical topological actions, i.e., representative grid reconfigurations.



Lehna et al. 2024, „Fault detection for agents on power gridtopology optimization: A comprehensive analysis."

Conclusion

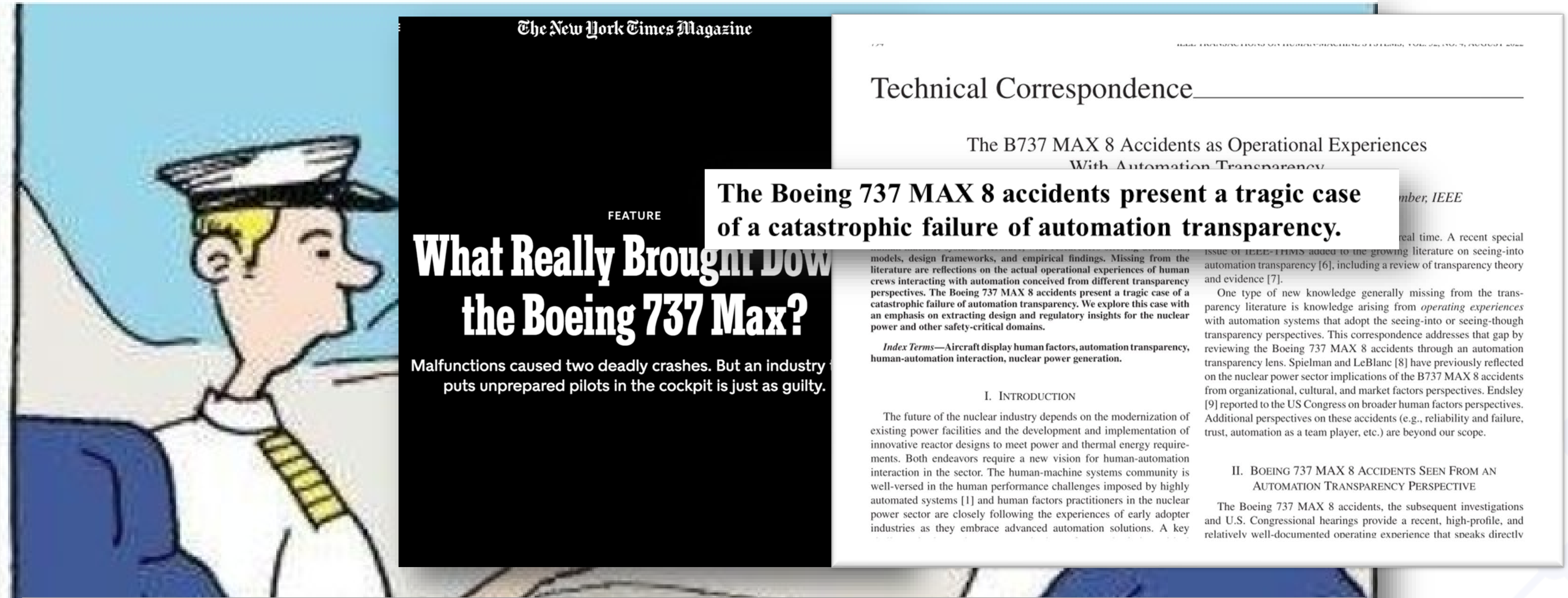
- Most research focuses on post-hoc explanation methods.
- However, post-hoc explanation methods are too unreliable for high-risk AI applications.
- More research on interpretable deep learning models is urgently needed.



Designing for Transparency

Clark Borst

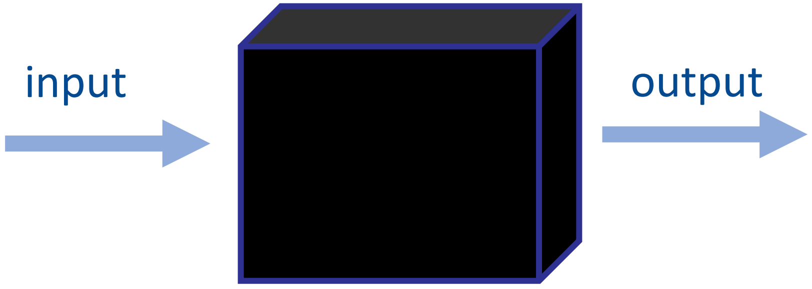
The importance of transparency



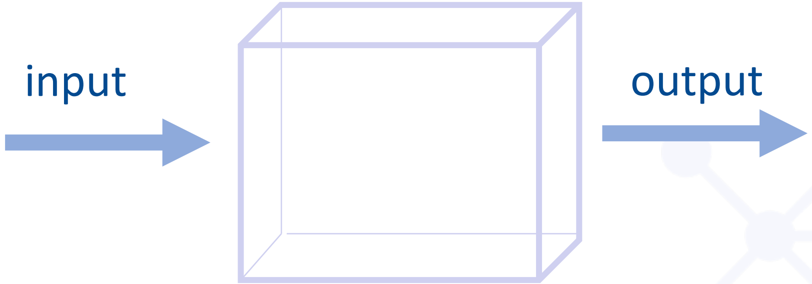
With increased automation, often less attention is paid to provide proper feedback to humans (because the role of humans seems less important), but the opposite is true!

Types of transparency

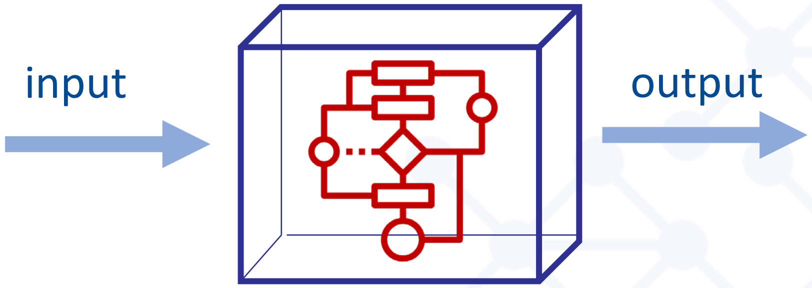
Black box automation: the human is deprived of knowledge and feedback about automation.



“seeing-through” transparency: create direct interaction between a human and automated task through a technology medium so well designed as to appear invisible.

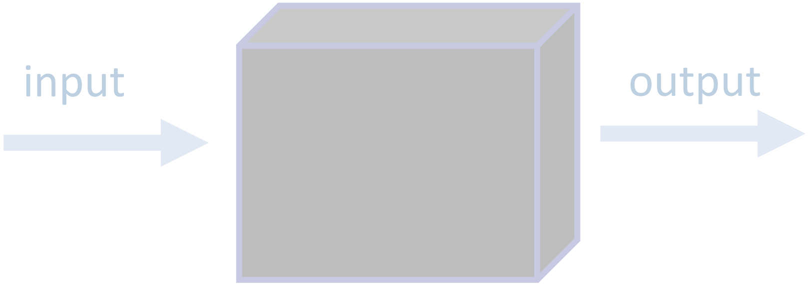


“seeing-into” transparency: facilitate human-automation interaction by revealing the automation’s responsibilities, capabilities, goals, activities, inner workings, performance, or effects to the human in real time.



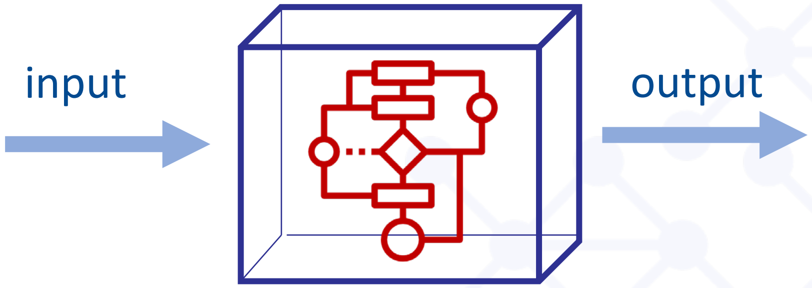
Types of transparency

Black box automation: the human is deprived of knowledge and feedback about automation.



“seeing-through” transparency: create direct interaction between a human and automated task through a technology medium so well designed as to appear invisible.

“seeing-into” transparency: facilitate human-automation interaction by revealing the automation’s responsibilities, capabilities, goals, activities, inner workings, performance, or effects to the human in real time.



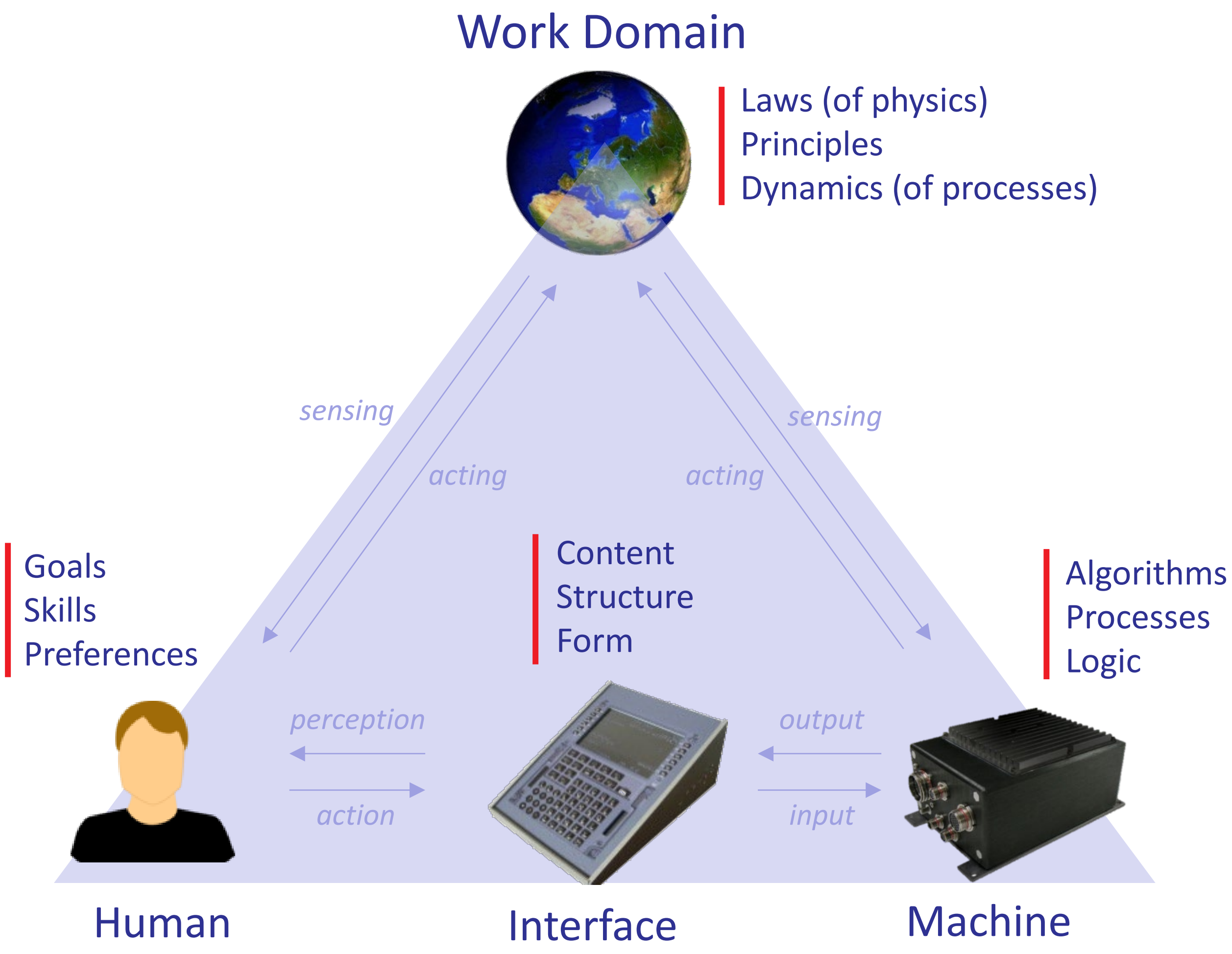
Designing for transparency

The **human-machine interface (HMI)** is a crucial element for successfully closing the (manual and supervisory) control loop.

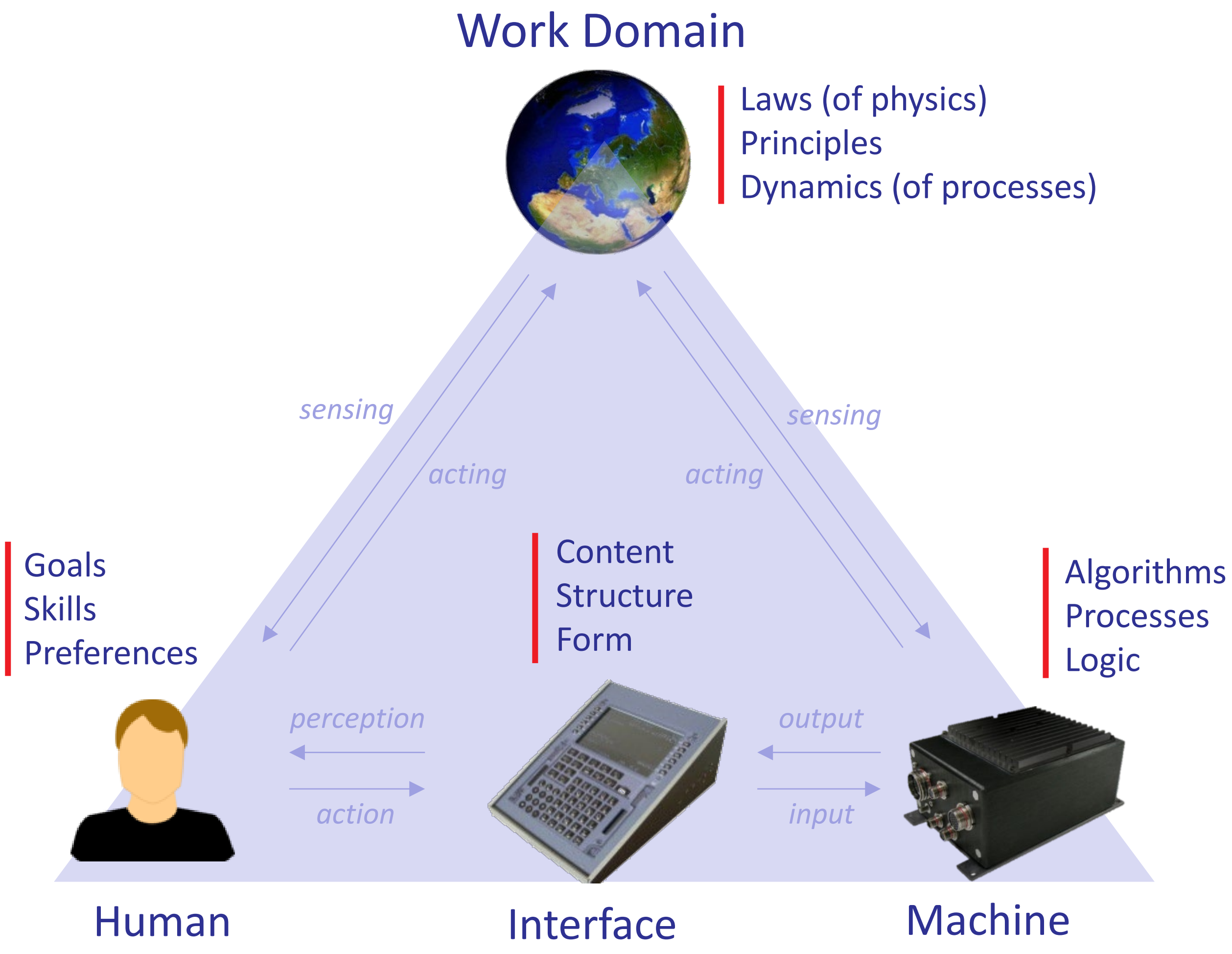
The goal of the interface is to “*provide the right information in the right way and at the right time.*” (Erik Hollnagel, 1988)

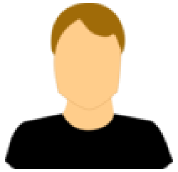
But, what is “*right*” and how can we find it?

Designing for transparency



Designing for transparency





What is the machine’s intent, solution and its achieved result (e.g., KPIs)?

user-centered *approaches, e.g., Situation Awareness-based Transparency (SAT) model*

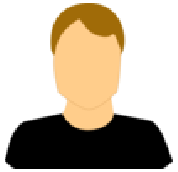
What physical and intentional constraints govern the machine’s solution(s)?

ecology-centered *approaches, e.g., Ecological Interface Design (EID)*

How does the machine explore the solution space? What does and doesn’t it consider?

model-centered *approaches, e.g., reward decomposition, search trees, decision trees, …*

Designing for transparency



What is the machine's intent, solution and its achieved result (e.g., KPIs)?

What physical and intentional constraints govern the machine's solution(s)?

How does the machine explore the solution space? What does and doesn't it consider?

Operational Transparency: how can an operational user be supported in understanding and assessing the (quality and validity of the) solution and operational impacts?

Domain transparency: what is the available solution space for human and automated agents to find solutions?

Engineering Transparency: how can a system developer be supported in designing and tuning the system?

Operational transparency

Interpretable Intentions

What goals are being pursued and what has priority?

Interpretable Solutions

Are solutions feasible and inline with domain requirements?

Interpretable Impacts

What are the impacts of the solutions on operational KPIs?

Engineering Transparency (XAI)

Explainable Data

What data was used to train and model and why?

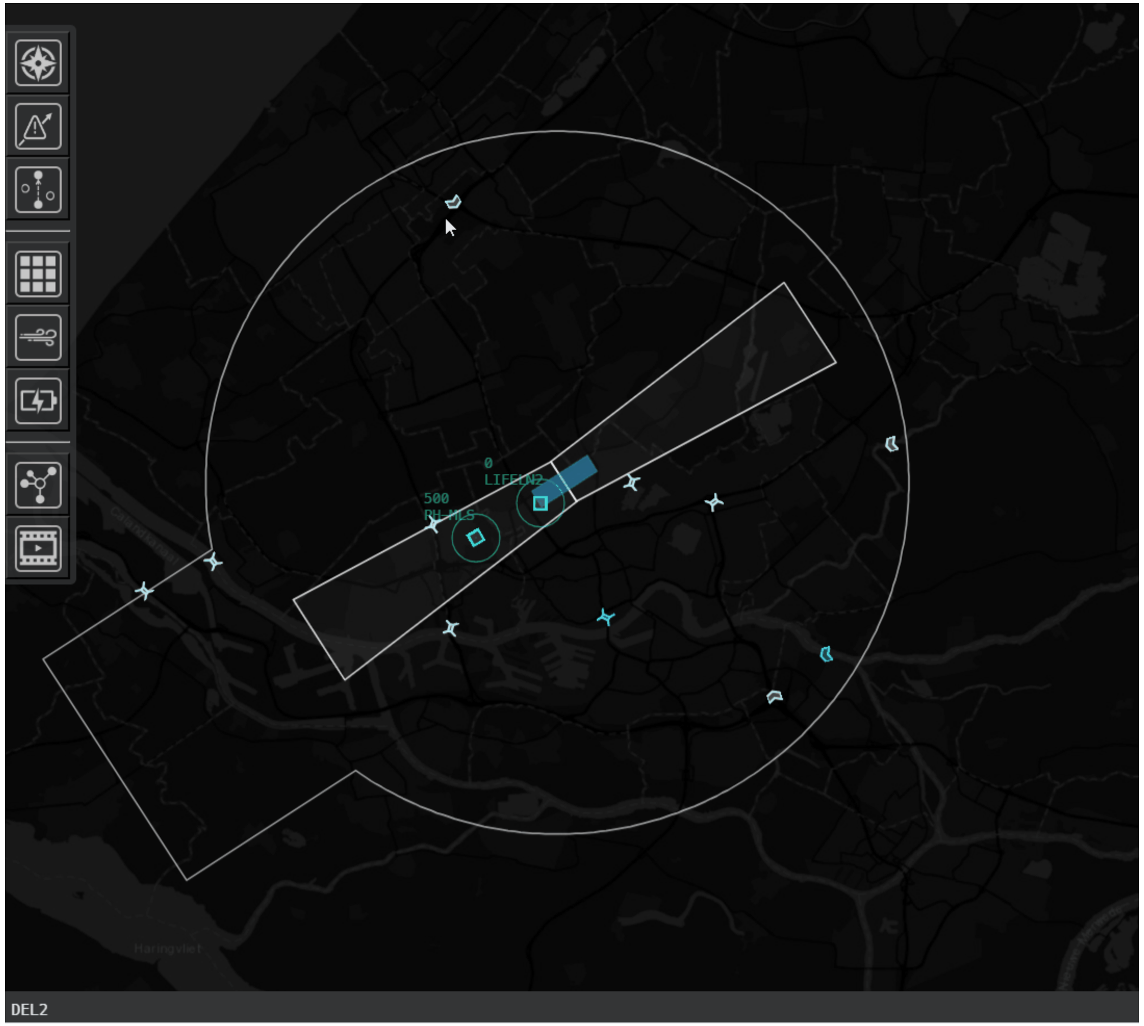
Explainable Predictions

What features and weights were used for this prediction?

Explainable Algorithms

What are the individual layers, thresholds, and logic used for a prediction or solution?

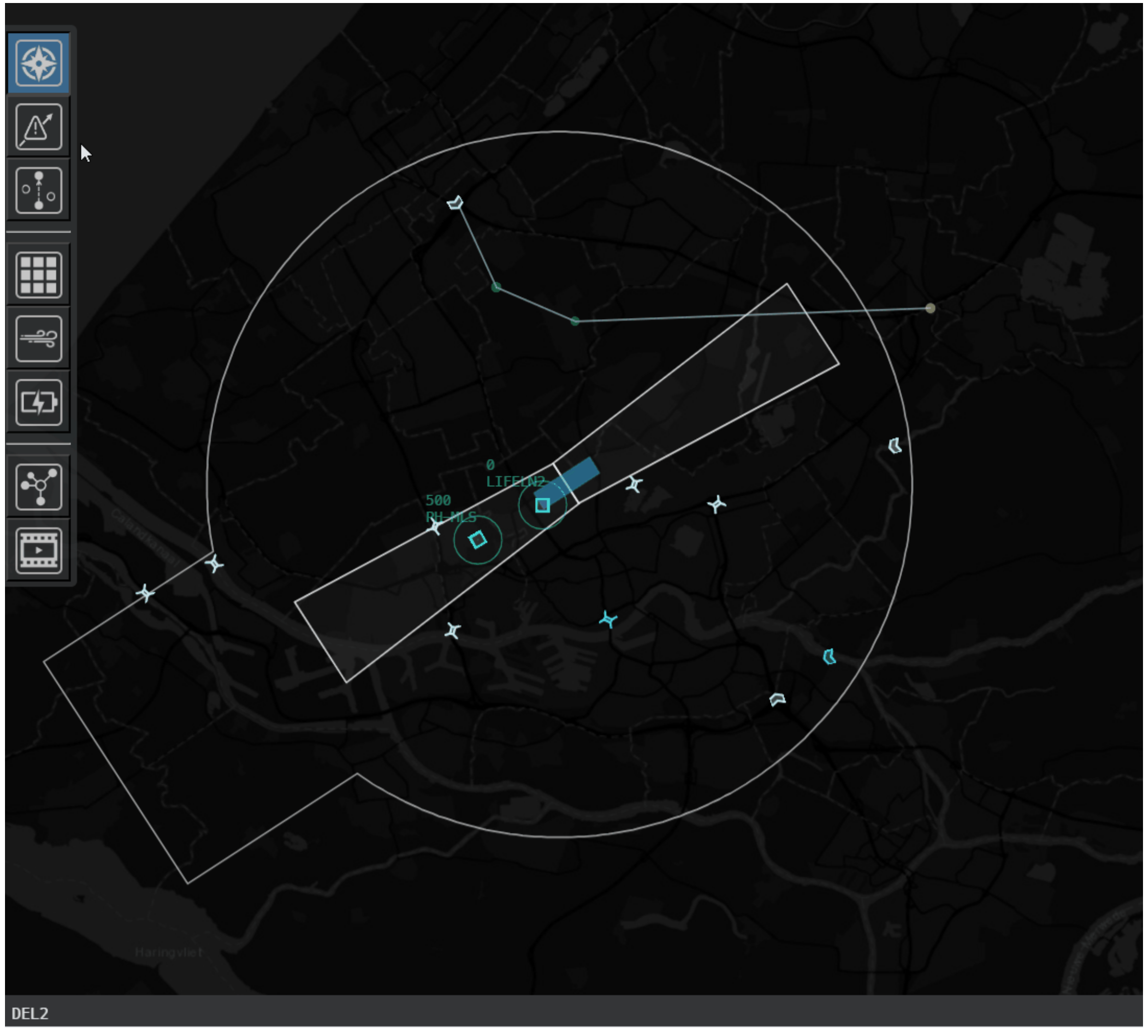
Example: transparency in drone pathfinding



What does the route look like?

What is the impact on the vehicle?

Example: transparency in drone pathfinding

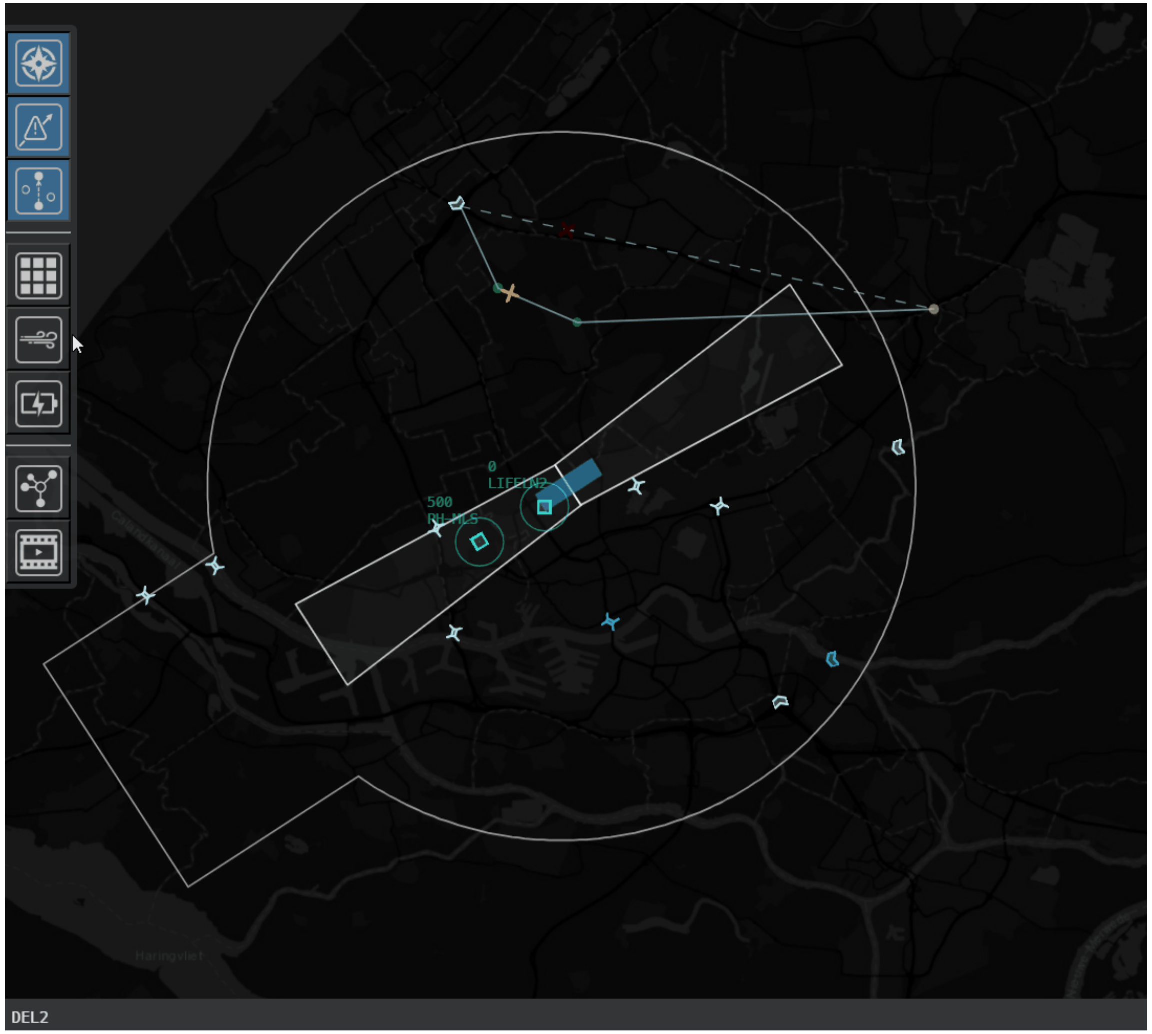


What does the route look like?

Is the route feasible and safe?

What is the impact on the vehicle?

Example: transparency in drone pathfinding



What does the route look like?

Is the route feasible and safe?

What is the impact on the vehicle?

What data is used?

Example: transparency in drone pathfinding

- What does the route look like?
- Is the route feasible and safe?
- What is the impact on the vehicle?
- What data is used?
- What settings were used to find the route?
- What options were explored and how?

Example: transparency in drone pathfinding

What does the route look like?

Is the route feasible and safe?

What is the impact on the vehicle?

What data is used?

What settings were used to find the route?

What options were explored and how?

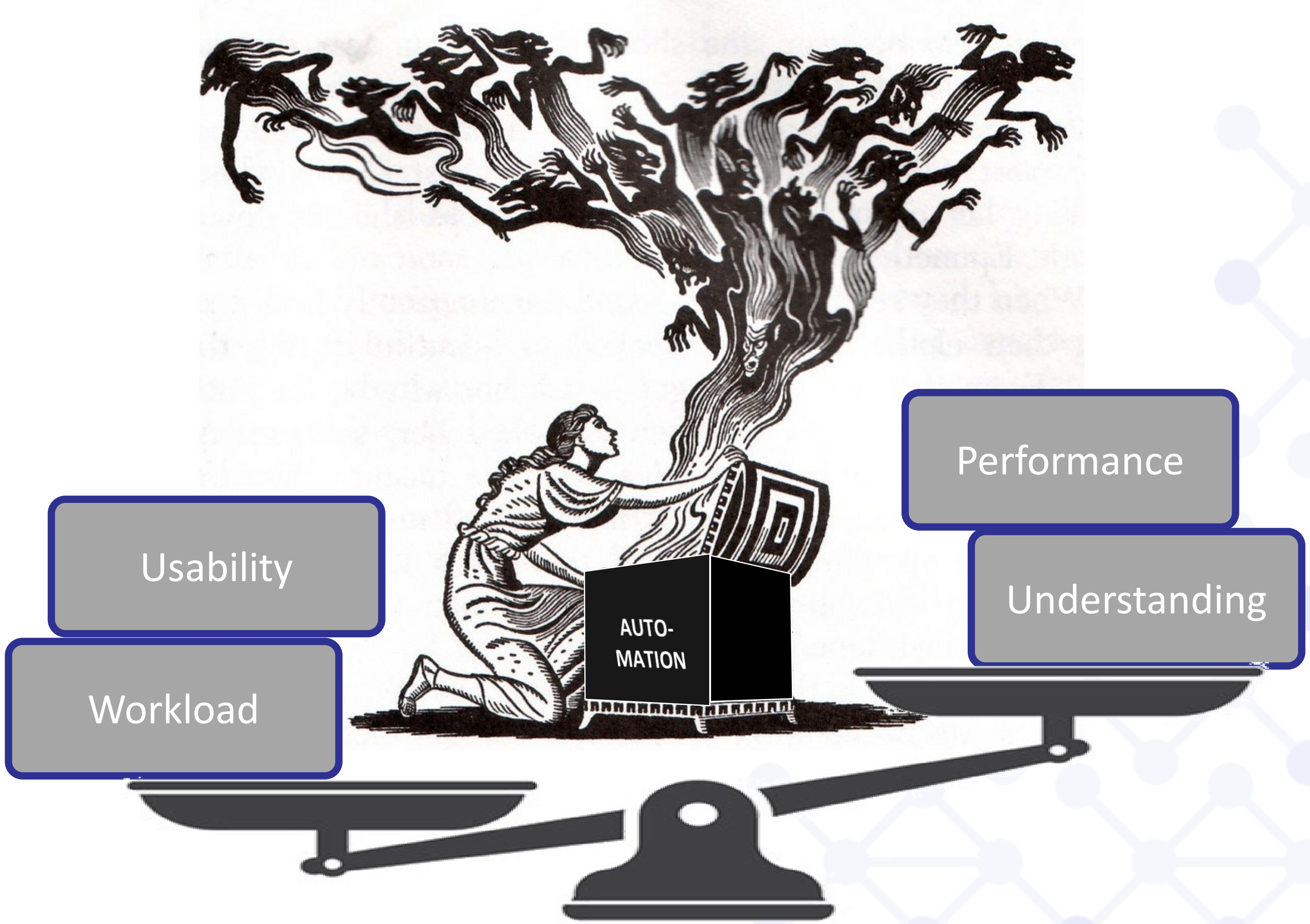
Transparency concerns

Opening the black box might unleash all sorts of evil (opening Pandora's Box):

Workload and usability
Increased transparency might come at the cost of increased workload (e.g., too much complexity) and usability issues (e.g., display clutter).

But..

Performance and understanding
Decreasing transparency might lower workload demands and improve usability, but may result in decreased understanding and reduced (supervisory control) performance.



Human Agency

Toni Wäfler

Human Agency

- Focus:
  - Operative decision-making
  - Experienced human experts and where the stakes are high

- Human agency
  - „… AI systems should support individuals in making better, more informed choices in accordance with their goals. …“ *

*European Commission: Directorate-General for Communications Networks, Content and Technology, *Ethics guidelines for trustworthy AI*, Publications Office, 2019

Problems from Perspective of Work Psychology

- Automation capabilities exceed human capabilities (Bainbridge, 1983)
  - Humans are assigned tasks that that go beyond their capabilities
- Automation complacency: Over-reliance (Parasuraman & Manzey, 2010)
  - Typical human errors: Omission error / commission error
- AI exacerbates these problems (Endsley, 2023)

Comprehensibility: Necessary but not Sufficient

- Humans still over-rely to AI even if the AI is comprehensible (by explainability / interpretability) (Buçinca et al., 2021; 2024)
  - Humans tend to not engage analytically with explanations
  - Cognitive forcing does not help
  - Humans tended to accept incorrect AI recommendations, even if they would have made a better decision without AI

- When humans do not engage with AI-generated functions and do not question them, performance decreases (Dell'Acqua et al., 2023)

From Recommendation-based AI to Supportive AI

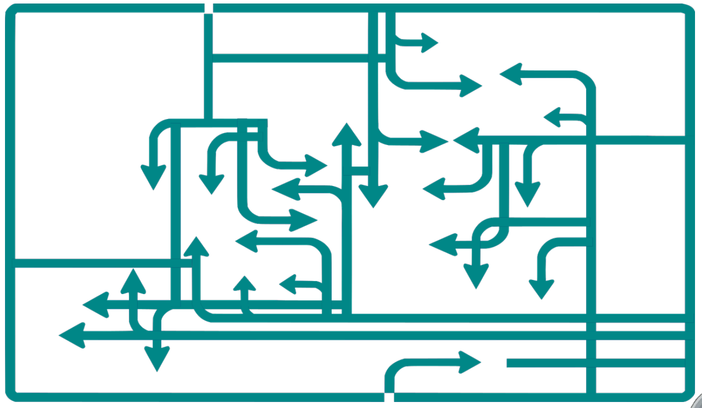


Image from Gerd Altmann auf Pixabay

Image from Mohamed Hassan auf Pixabay

Recommendation-based AI
- Sophisticated recommendations
- Over-reliance of humans

Supportive AI
- Supporting cognitive processes
- Augmenting human cognition

Image from OpenClipart-Vectors auf Pixabay

Supportive-AI Explicitly Supports Human Cognitive Processes

- Supporting human decision-making regarding:
  - e.g. managing attention
  - e.g. comparing effects of different options for decisions
- Supporting human learning regarding:
  - e.g. building expertise regarding leverage points
  - e.g. identifying weak signals of emerging problems
- Supporting human motivation by:
  - e.g. making transparent causal relations (for experienced meaningfulness)
  - e.g. providing feedback regarding the impact of their decisions (for feedback)

AI4REALNET has received funding from European Union's Horizon Europe Research and Innovation programme under the Grant Agreement No 101119527

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. Neither the European Union nor the granting authority can be held responsible for them.