

Trustworthy AI-Native 6G: Explainable and Robust Intelligence for Resilient and Secure PHY Layer

Abdulkadir Çelik

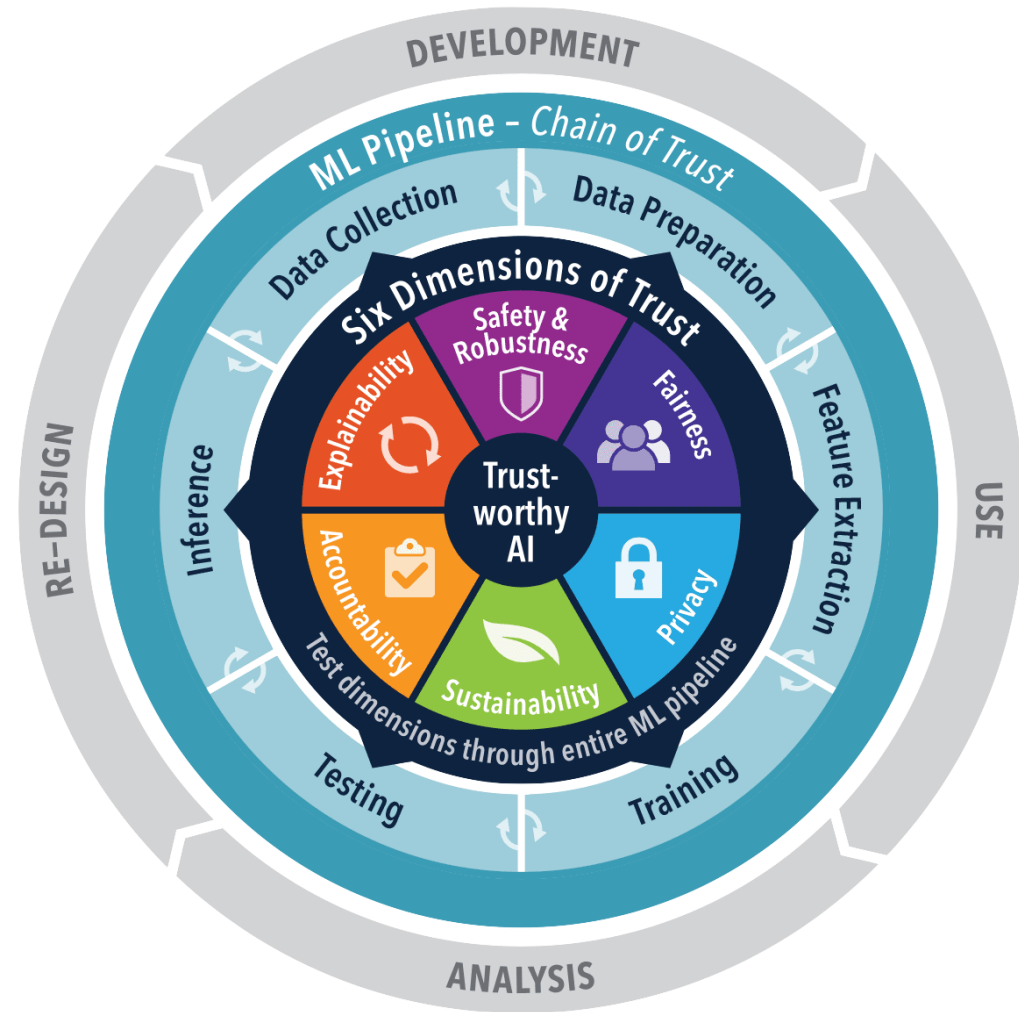
Associate Professor
Head of Centre for IoT and Pervasive Systems
Director of Wireless Intelligence & Networked Sensing Lab
School of Electronics and Computer Science
University of Southampton, UK



**WIRELESS WORLD
RESEARCH FORUM
HUDDLE**

*One Vision
Many Voices
16-17 September
Istanbul, Türkiye*

Six Dimensions of Trustworthy AI



Robustness & Resilience

Reliable behaviour under noise, rare events, and attack.



Explainability

Decisions a human can inspect, follow and challenge.



Privacy

Personal data protected across training and inference.



Sustainability

Compute, energy and lifecycle cost kept proportionate.



Accountability

Clear ownership, audit trail and recourse for decisions.

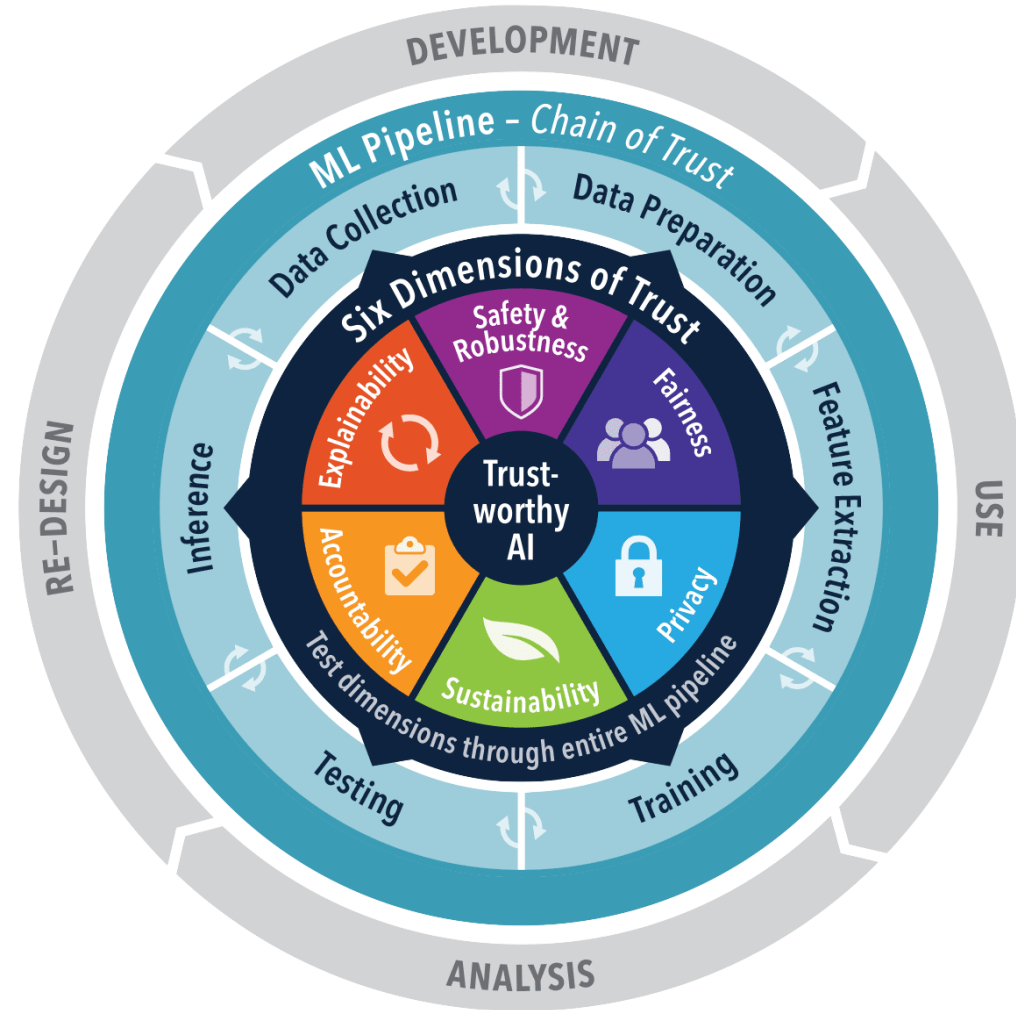


Fairness

No systematic disadvantage across groups or subpopulations.

*EU: AI Act & GDPR,
US: Safe, Secure & Trustworthy AI*

Six Dimensions of Trustworthy AI



Robustness & Resilience

Reliable behaviour under noise, rare events, and attack.



Explainability

Decisions a human can inspect, follow and challenge.



Privacy

Personal data protected across training and inference.



Sustainability

Compute, energy and lifecycle cost kept proportionate.



Accountability

Clear ownership, audit trail and recourse for decisions.

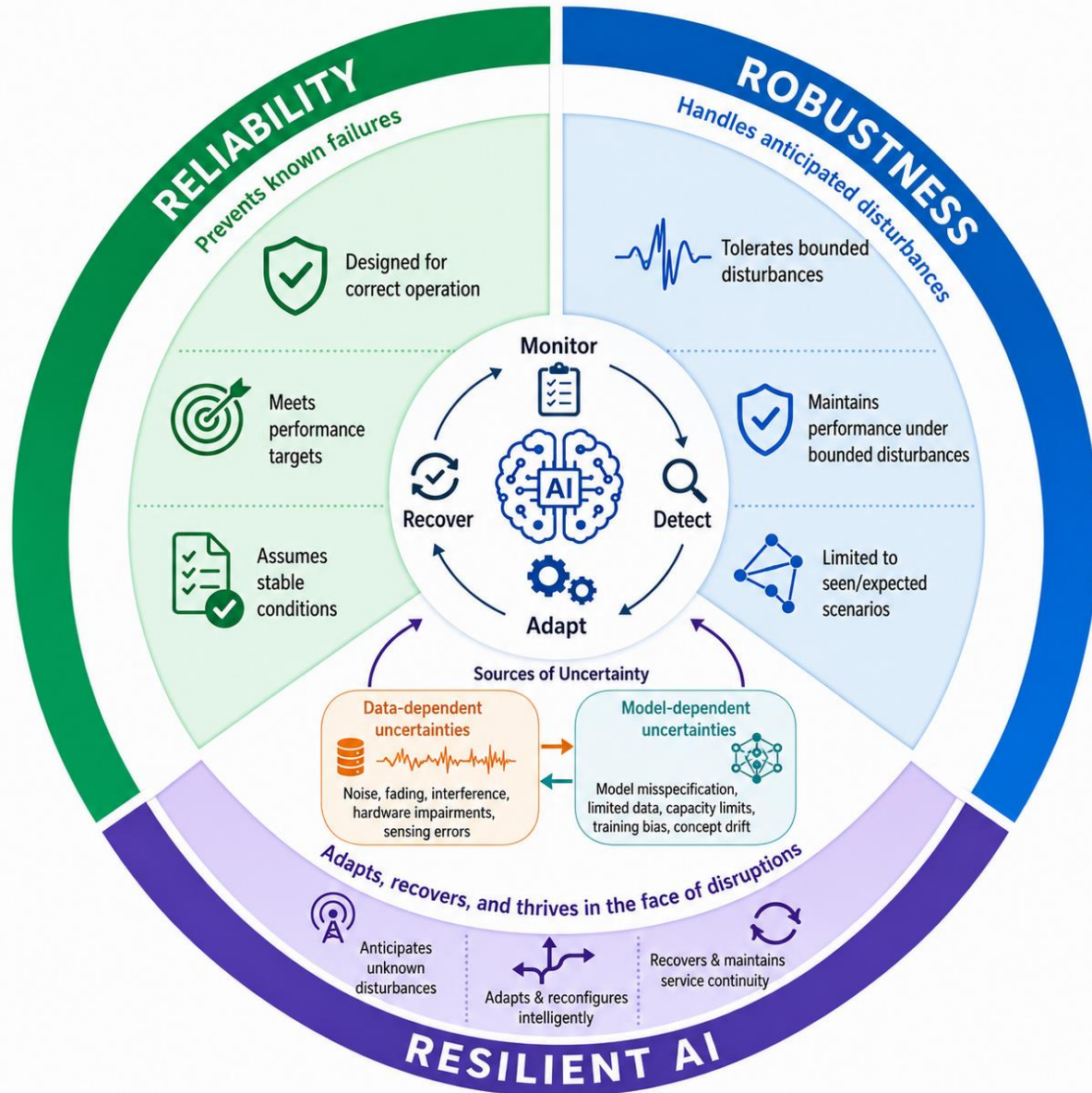


Fairness

No systematic disadvantage across groups or subpopulations.

*EU: AI Act & GDPR,
US: Safe, Secure & Trustworthy AI*

Robustness buys tolerance; resilience buys recovery speed



Reliability

Meets predefined KPIs (outage, latency, error, rate) under assumed conditions.

Robustness

Tolerates bounded uncertainty with no adaptation, retraining or policy change.

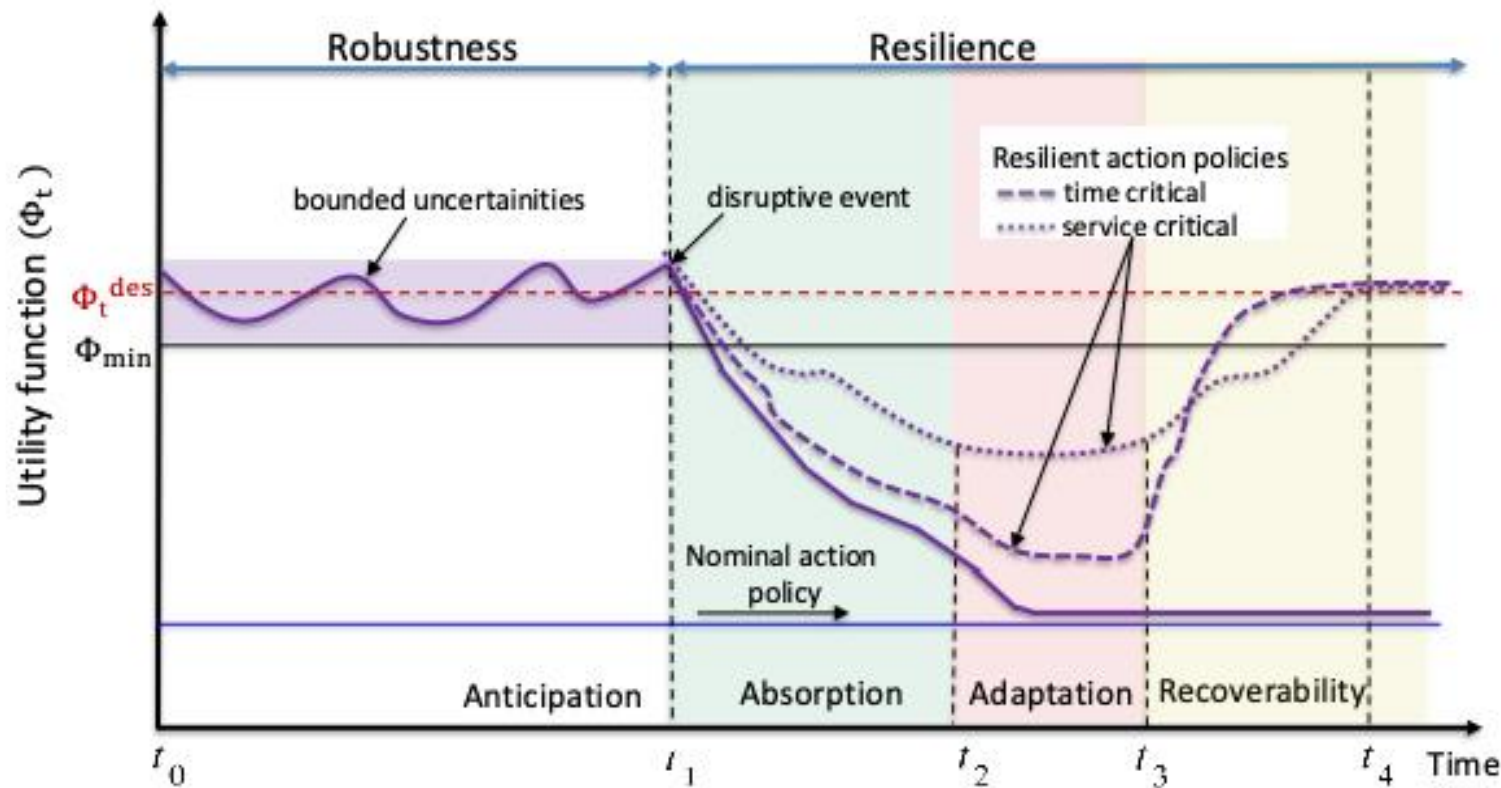
C Resilience

Absorbs, adapts and restores utility after disruptions outside the training envelope.

Recovery latency, not tolerance, decides the outcome after t_1



Utility function $\Phi(t)$ against time, one disruptive event, three control policies



The tradeoff: the time-critical policy dips deeper but recovers fast; the service-critical policy protects utility during the event but recovers slowly. The nominal policy never recovers — it falls through Φ_{min} and stays there.

The four resilience phases

- **Anticipation:** spot the precursors and act before t_1 .
- **Absorption:** how much utility drop the model handles on its own
- **Adaptation:** change the policy to arrest the decline and claw utility back.
- **Recoverability:** converge to a stable post-event operating point.

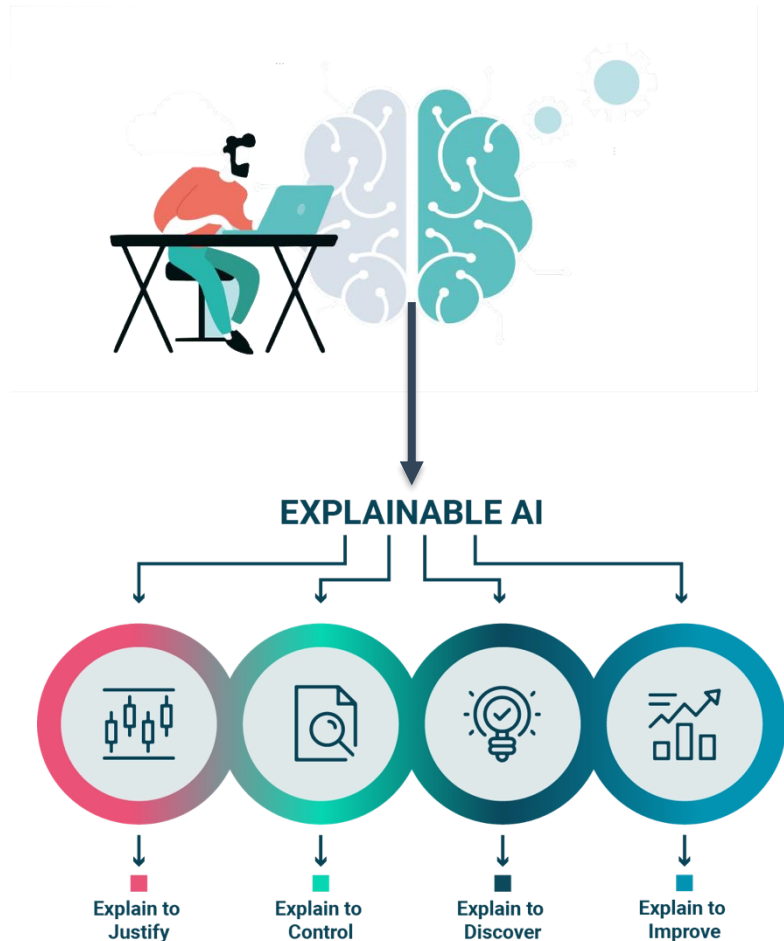
Retraining sets the clock

- $R_L = t_4 - t_1$ is set by communication and compute: edge/cloud capacity must adapt and run the new policy inside the TTI budget.
- **Full retraining rarely fits: favour online learning and pre-computed fallbacks.**

Explainable AI serves four different purposes



Explainable AI is not one technique. It is four different jobs, and which one you need decides what a good explanation looks like.



Explain to Justify

Defend a decision after the fact to a regulator, an auditor, or the person it affected.

Explain to Control

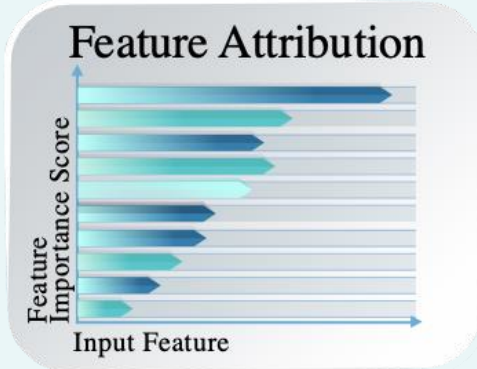
Catch faulty reasoning before it acts: verify the model relies on the right evidence.

Explain to Discover

Learn something new from the model such that relationships in the data nobody had noticed.

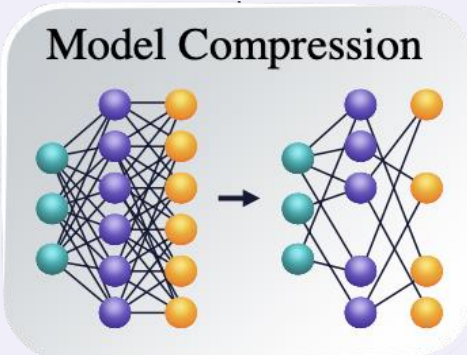
Explain to Improve

Use the explanation to debug and refine the model itself, not just describe it.



Feature Attribution

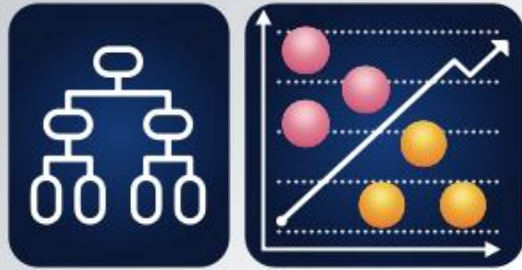
- Generate human-readable interpretations
- Provide meaningful insights about the model behavior
- Provide clear and quantifiable explanations for feature importance
- SHAP (SHapley Additive exPlanations) for feature rankings



Model Compression

- Reduce size and complexity of models while preserving accuracy
- Reduce energy and memory needs for resource-constrained systems
- Reduce control & communication overhead, boost convergence speeds

Model Simplification

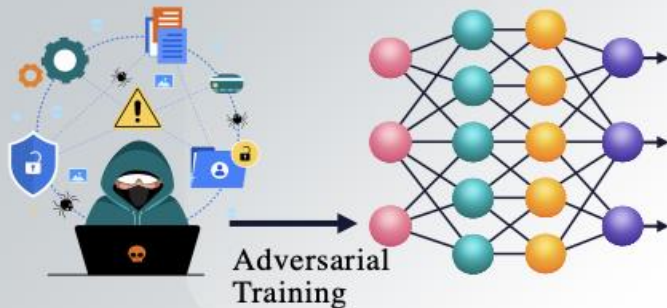


Decision Tree Linear Regression

Model Simplification

- Surrogate models to simplify the perception of the complex ones
- Creates locally faithful representation of original model behavior
- Control trade-off between explanation accuracy and interpretability
- Local Interpretable Model-agnostic Explanations (LIME) and knowledge distillation

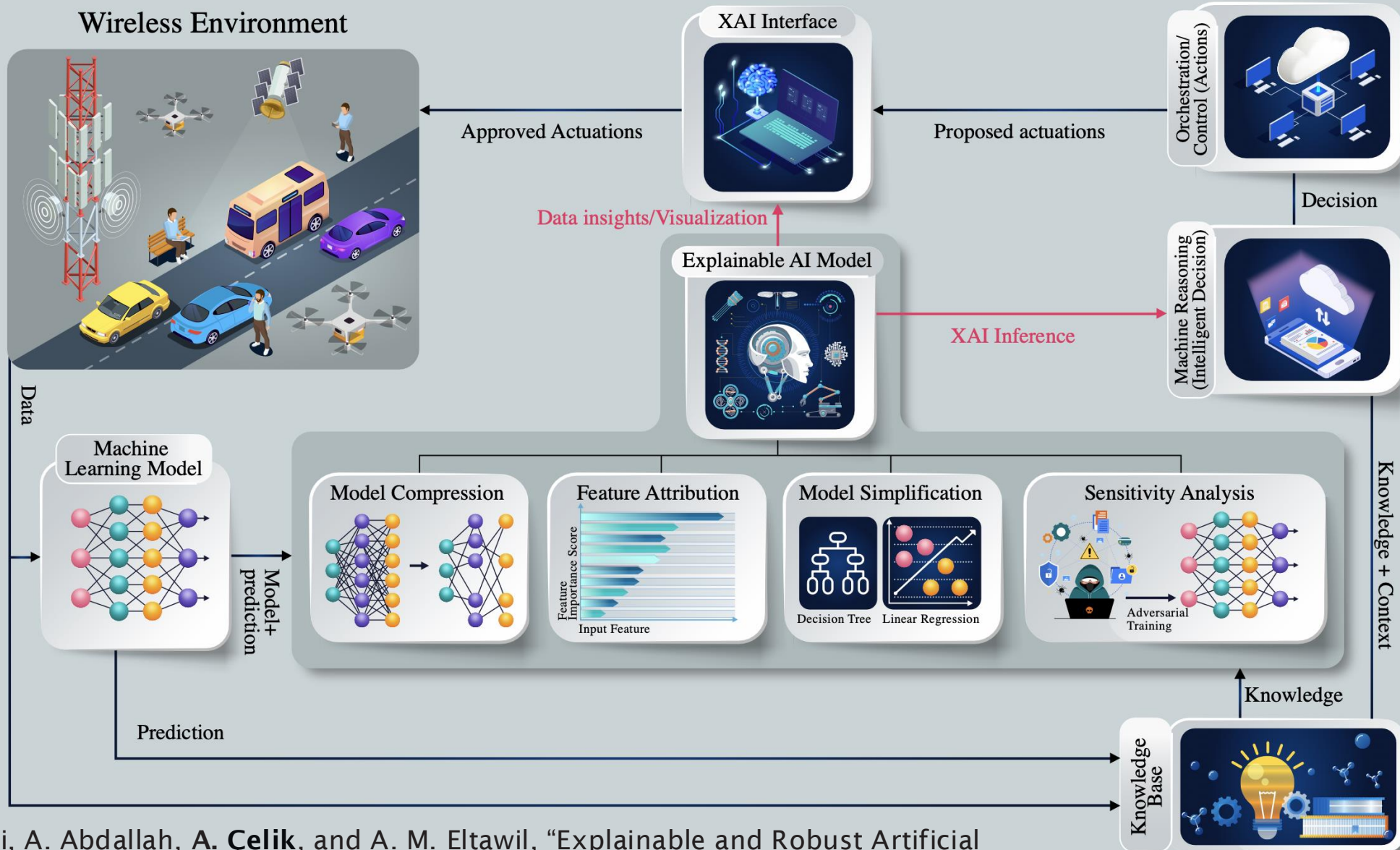
Sensitivity Analysis



Sensitivity Analysis

- Identifying potential outliers and malicious input
- Quantifying the model output's sensitivity to changes in the input data
- Resilience against adversarial attacks

Explainable and Robust RRM

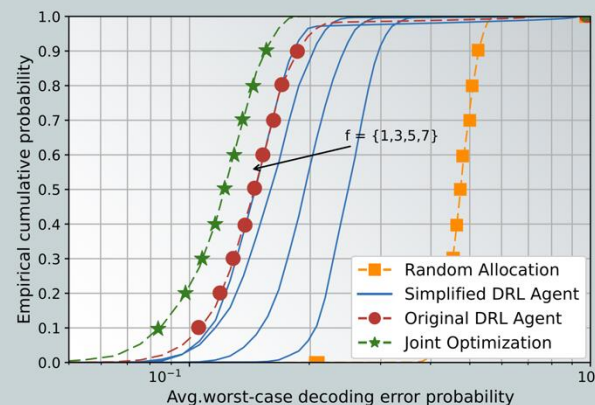
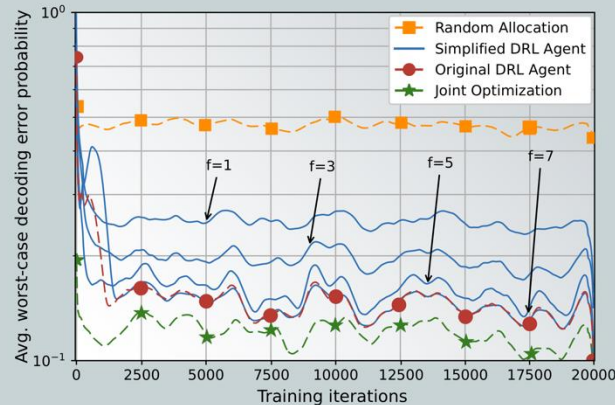
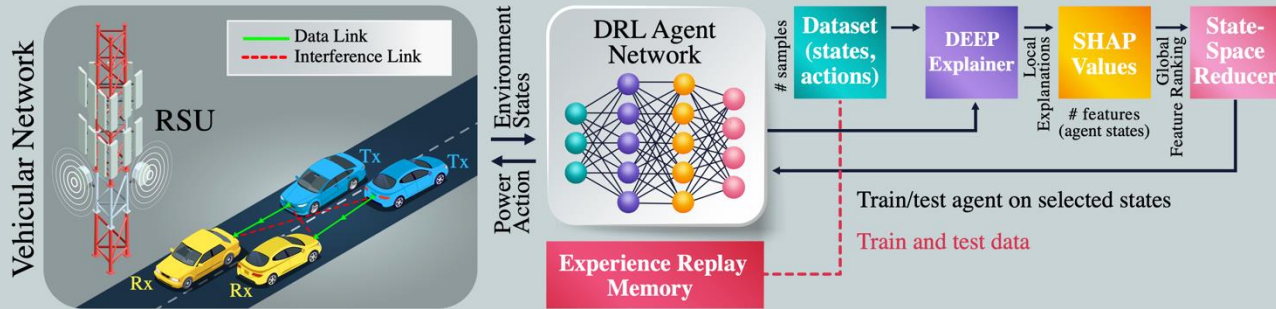


Case I: DRL-Based RRM in V2X Network



DRL based model training & execution

Post-hoc explainability pipeline

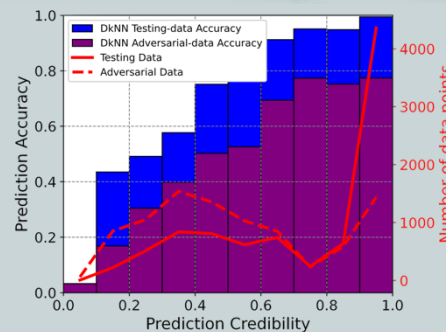
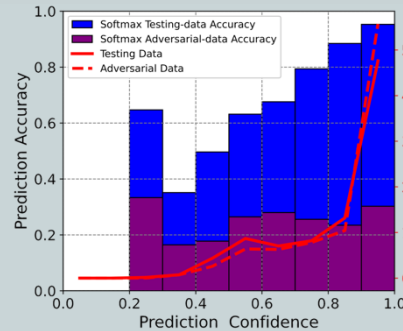
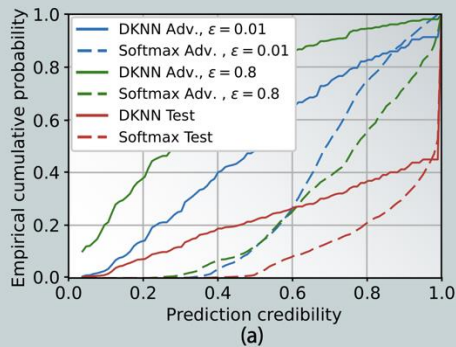
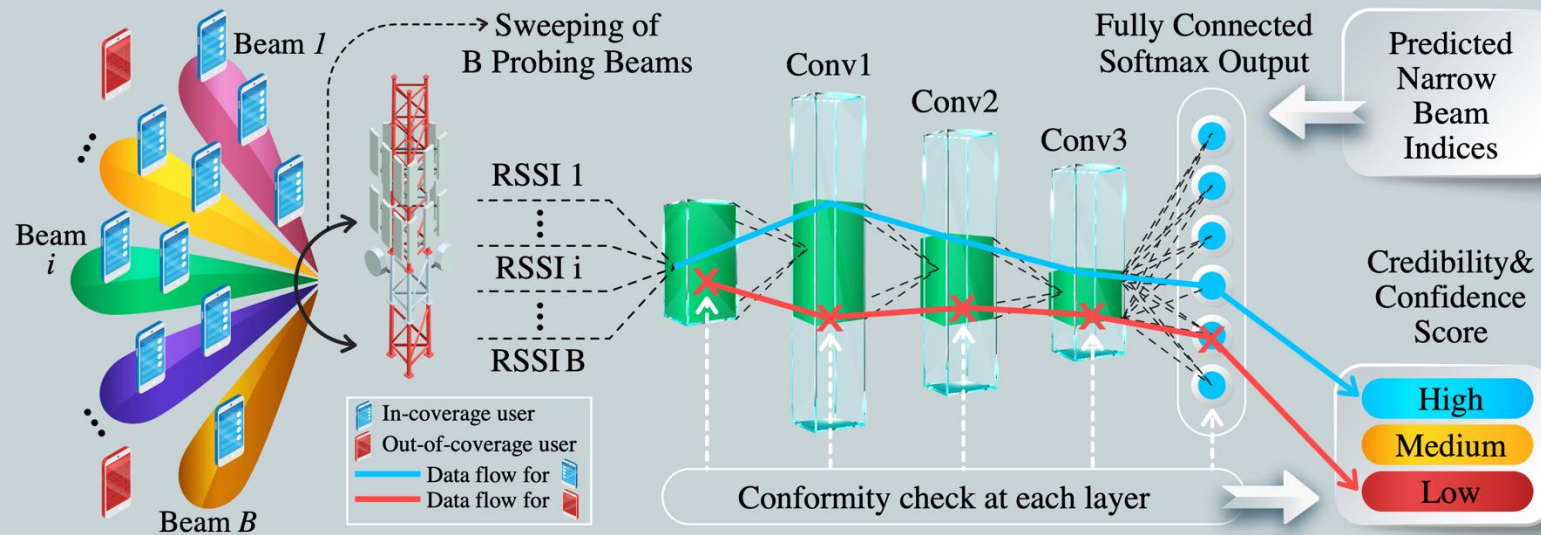


- Complexity increases with the size of the state space cardinality ($2K+2$)
- XAI-based state importance ranking
- Performance vs. simplification trade-off

Model simplification [$f=7$] reduces

- # Input states by **~68%**
- # Broadcasted training param. by **~81.5%**
- Average training time by **~49.1%**
- Performance degradation **<1%**

Case I: DRL-Based RRM in V2X Network



Conformity check via Deep-KNN

- Generate the out-of-training (adversarial) data
- Softmax overestimates the confidence
- Deep-kNN measures the discrepancy between labeled inputs and past observations

- ✓ High accuracy with high credibility on majority test data points
- ✓ Low prediction credibility on majority of the adversarial points (outliers)
- ✓ Deep-kNN assigns low credibility (<0.5) to outliers, indicating a lack of support from the training data

YOUR QUESTIONS

Abdulkadir Celik

e: a.celik@soton.ac.uk

e: c-iot-hog@ecs.soton.ac.uk

w: www.c-iot.ecs.soton.ac.uk

