

Overview of Methodologies for Testing AI Robustness and Explainability in Highly Dynamic and Scalable Environments

Thanasis Kotsiopoulos, Alexandros Nizamis,
Usman Wajid, Konstantinos Votis



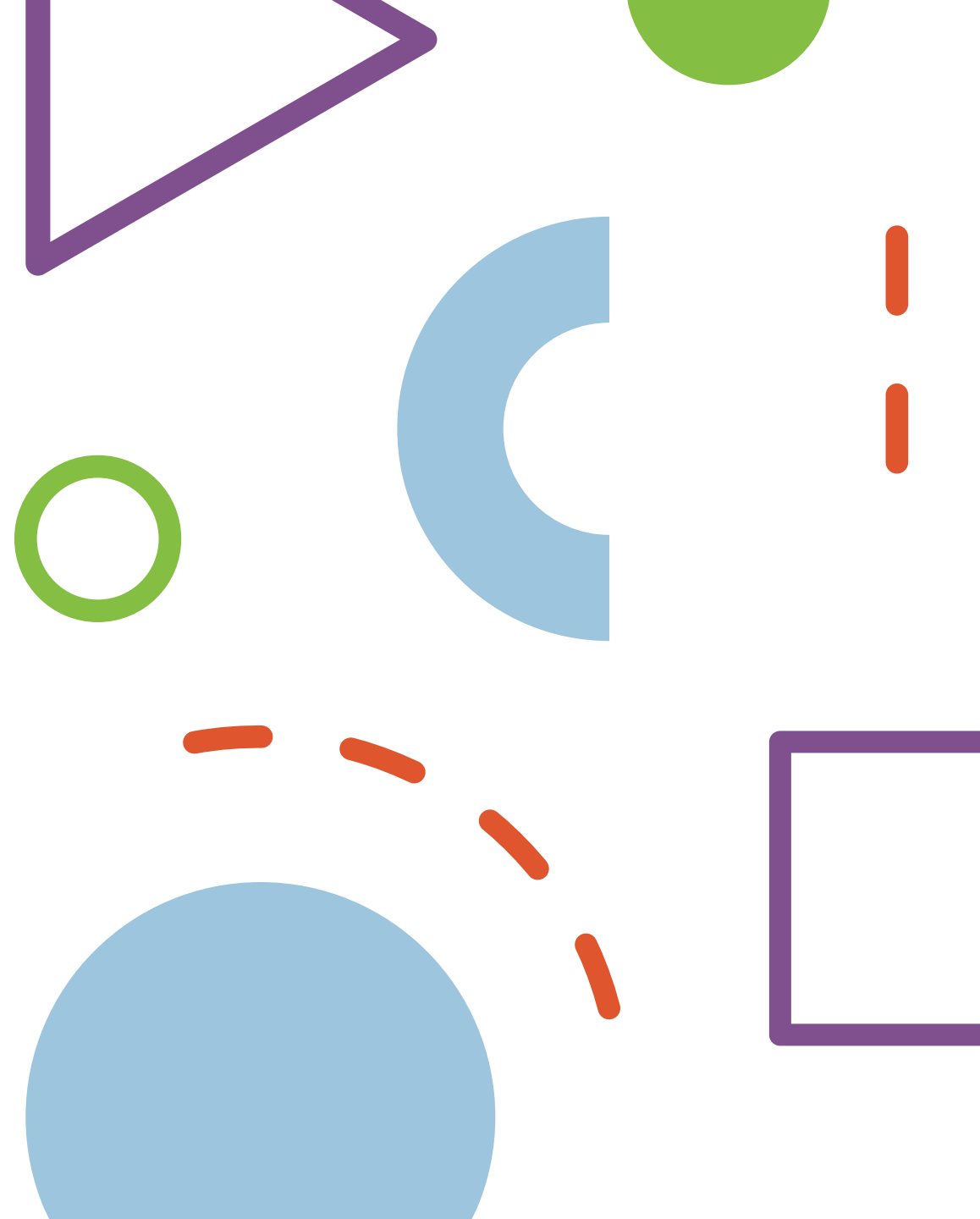
CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

02/04/2025



Agenda

- Introduction & Problem Statement
- Industrial Context & Motivation
- Testing Methodologies
- Technological Implementation
- Integrating X-AI
- Outcomes & Lessons Learnt
- Future Directions
- Q&A



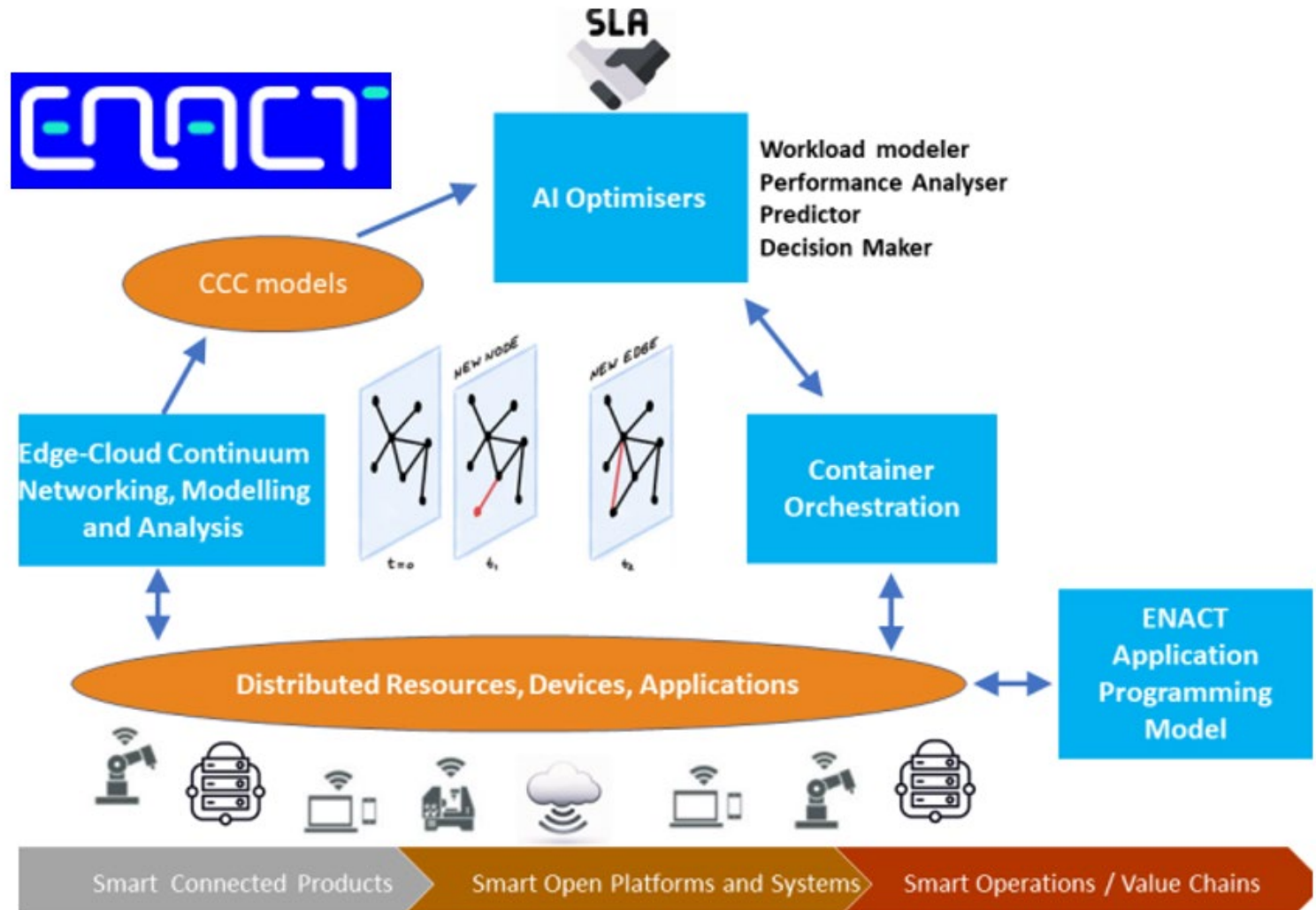
What is ENACT



Tools to connect and discover distributed resources, devices and services across the compute continuum, characterize and model them to support complex application deployment needs

AI-powered orchestrator capable of deploying and managing applications across distributed (edge - cloud) nodes in an optimal way to support the energy efficiency and adaptations in applications

Application development toolkit for developing or adapting complex applications, making them distributed, responsive, robust and adaptive to changing environments



Problem Statement

Challenge

Ensuring robustness and explainability of AI models amid dynamic, distributed edge-cloud environments

Problem Context

1. Traditional testing methods do not capture real-time variability and concept drift
2. AI models require continuous monitoring and transparent evaluation

Importance

1. Minimizing risk,
2. maintaining performance
3. enabling adaptive responses

Industrial Context & Motivation

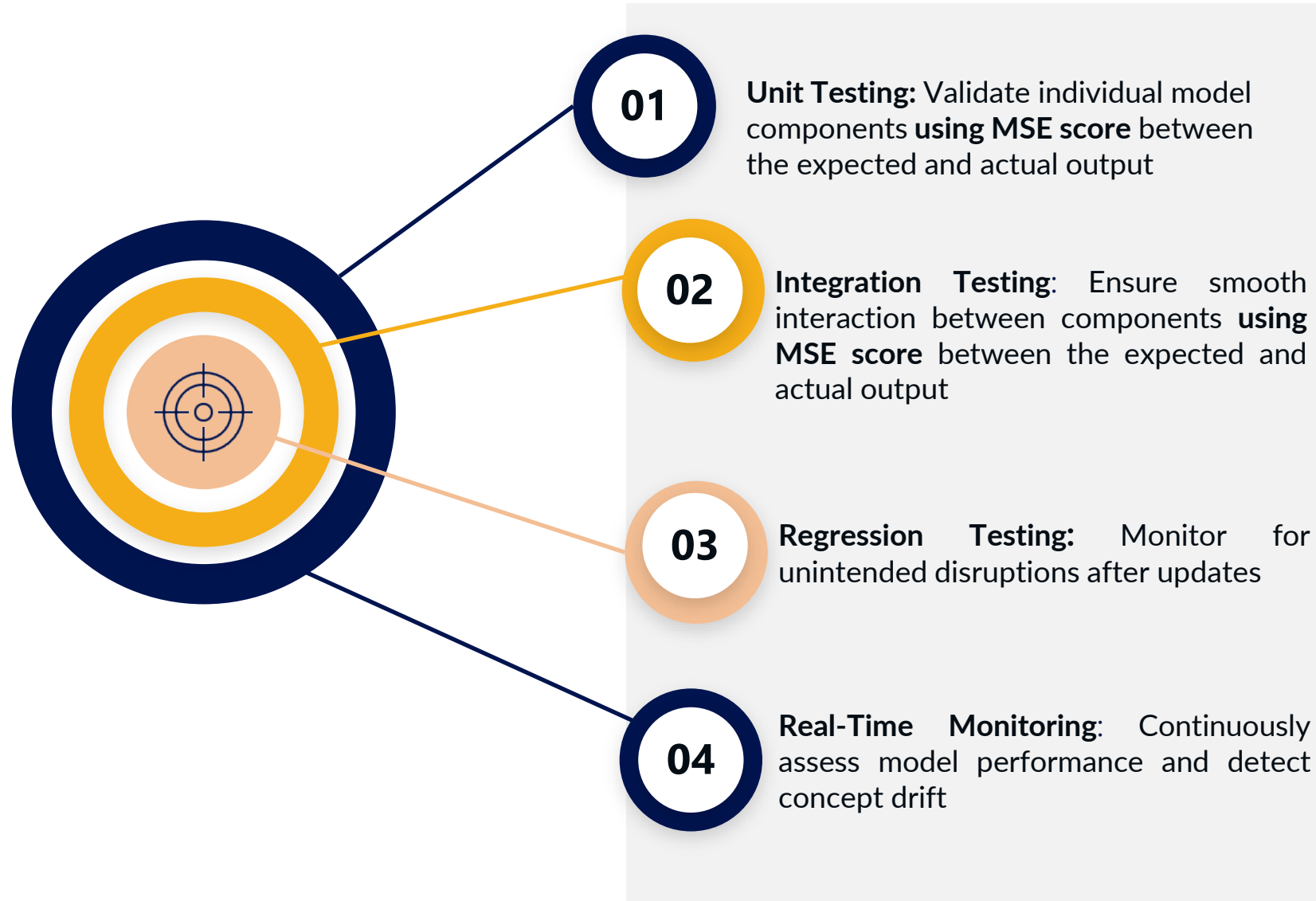
Relevance to Industry

- Automated testing is key to industrial deployment and CI/CD pipelines
- The ENACT project drives orchestration solutions for highly distributed applications

Motivation

- Bridging the gap between research and practical applications
- Ensuring that testing frameworks evolve with dynamic environments

Testing Methodologies Overview



Technological Implementation

- CI/CD pipeline for automated deployment and validation
- ENACT's AI Compliance Checker integrated with XAI for detailed insights
- Automated real-time model monitoring systems
- Data-driven decision-making via continuous monitoring

```
1 stages:
2   - build
3   - train
4   - xai
5   - test
6   - deploy
7
8 # Stage 1: Build - install dependencies and set up the environment
9 build:
10  stage: build
11  image: python:3.10
12  script:
13    - echo "Installing dependencies..."
14    - pip install -r requirements.txt
15  artifacts:
16    paths:
17      - .venv/
18
19 # Stage 2: Train Model - run the training script for the deep reinforced learning agent
20 train_model:
21  stage: train
22  image: python:3.10
23  script:
24    - echo "Training the deep reinforced learning agent..."
25    - python train_agent.py
26  artifacts:
27    paths:
28      - models/
29
30 # Stage 3: XAI Analysis
31 xai_analysis:
32  stage: xai
33  image: python:3.10
34  script:
35    - echo "Running XAI analysis..."
36    - python xai_analysis.py
37  artifacts:
38    paths:
39      - xai_reports/ # store explanation report
40
41 # Stage 4: Test - run unit and integration tests, including tests that check the validity of the XAI outputs
42 test:
43  stage: test
44  image: python:3.10
45  script:
46    - echo "Running tests unit and integration tests"
47    - python run_tests.py
48
49 # Stage 5: Deploy - deploy the model and related artifacts if tests pass
50 deploy:
51  stage: deploy
52  image: python:3.10
53  script:
54    - echo "Deploying the model..."
55    - python deploy_model.py
56  only:
57    - dev # deploy only from the dev branch
58
```

Integrating X-AI - Overview

Techniques Introduced

1. Integrated Gradients
2. LIME
3. SHAP values
4. ENACT's AI Compliance Checker

Purpose of XAI Integration

1. Enhance transparency by providing feature-level insights
2. Aid risk assessment and uncover biases

Outcome

1. Enables early detection of performance shifts and builds trust in AI models

Integrating X-AI – Test Case 1

- Objective: Demonstrate that features with high attribution scores are crucial to the AI models' performance.
- Methodology:
 - Identify high-attribution features from the Integrated Gradients/LIME/SHAP output.
 - Mask these features from the input and observe the agent's decision or performance
 - Compare the performance degradation with scenarios where low-attribution features are removed and raise assertion errors if accuracy is deviating over 10%.
- Expected Outcome: Removing high-attribution features should significantly reduce the agent's performance, validating that the explanation accurately reflects the model's reliance on those inputs.



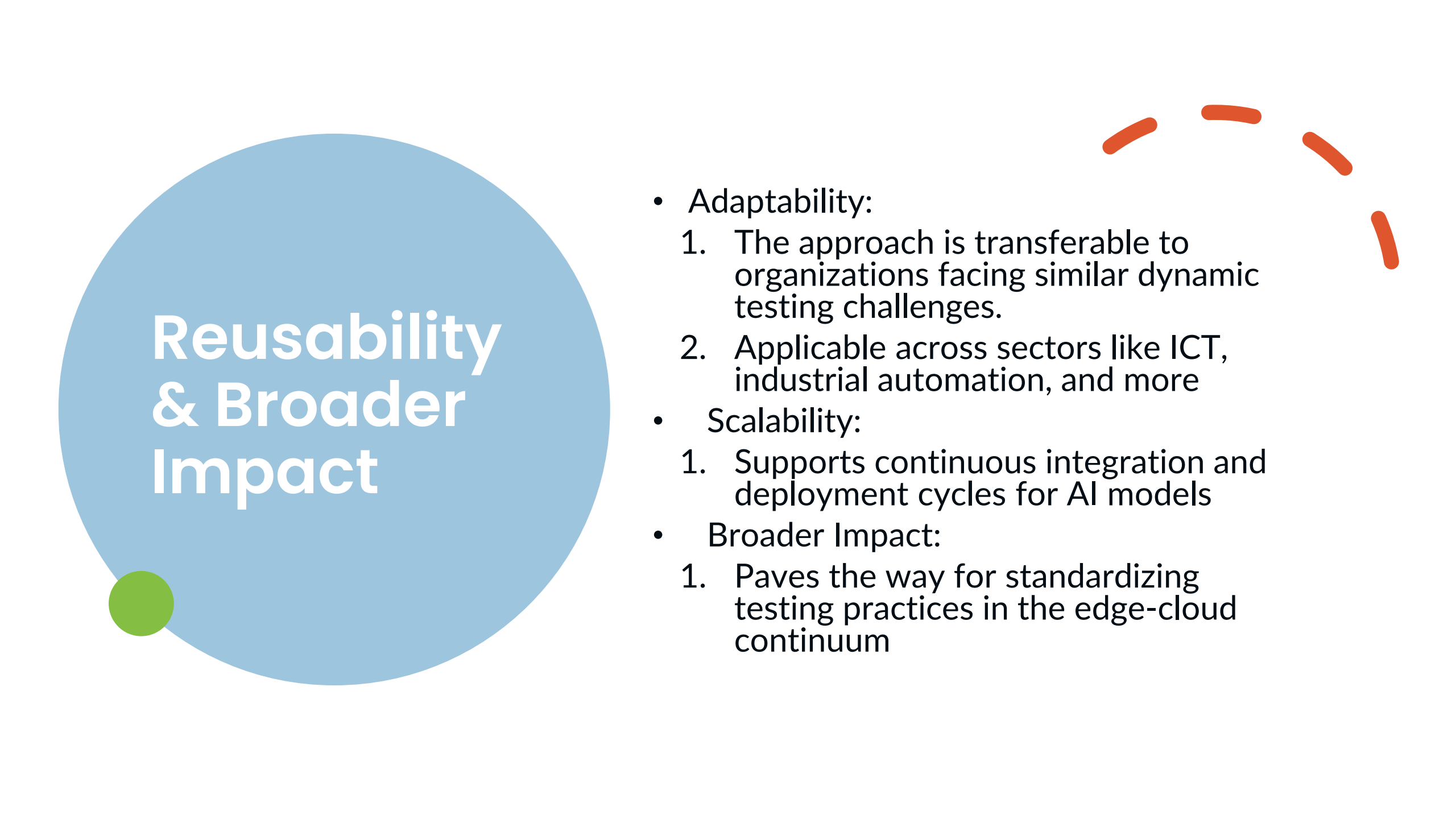
Integrating X-AI – Test Case 2

- Objective: Demonstrate that AI libraries and methodologies used are compliant with AI best practices
- Methodology:
 - Run the AI Act Compliance Checker.
 - Obtain the issues.
 - If there are at least two issues of high severity, raise an assertion error
- Expected Outcome: Apply strict policy on how AI is developed using standardized procedures.



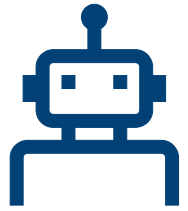
Outcomes & Lessons Learned

- **Key Outcomes:**
 1. Enhanced detection of concept drift and dynamic performance shifts
 2. Improved transparency of AI decision-making processes
- **Lessons Learned:**
 1. Integration of XAI within testing frameworks is promising but a lot of calibration is needed when new datasets are introduced.
 2. Balancing automated testing with explainability builds more trust between human and AI collaboration
- **Implication:**
 1. Provides a scalable approach that can be adopted by various industries



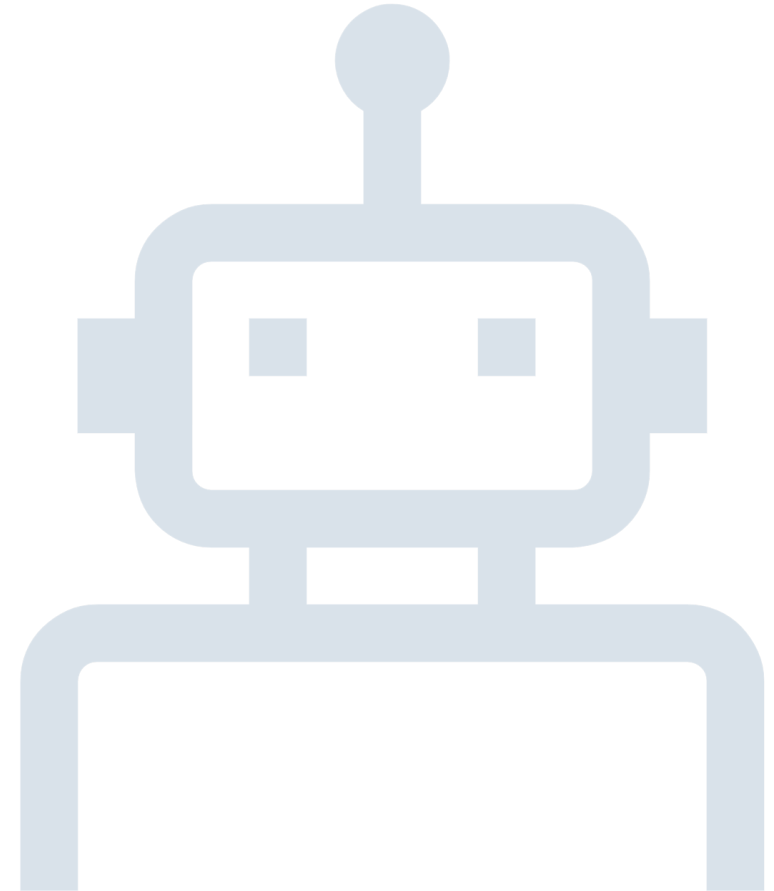
Reusability & Broader Impact

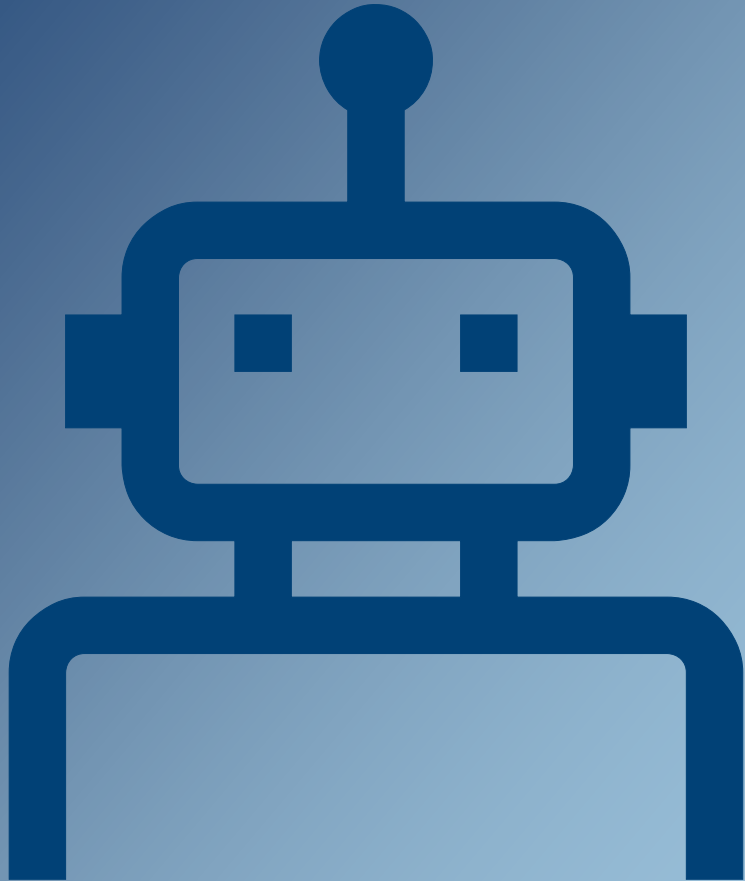
- Adaptability:
 1. The approach is transferable to organizations facing similar dynamic testing challenges.
 2. Applicable across sectors like ICT, industrial automation, and more
- Scalability:
 1. Supports continuous integration and deployment cycles for AI models
- Broader Impact:
 1. Paves the way for standardizing testing practices in the edge-cloud continuum



Future Directions & Next Steps

- Research & Development:
 - Scale the testing framework for larger, more complex deployments
 - Partnerships with industry for real-world validations
 - Continuous update of the AI Act Compliance Checker and integration with AI emerging standards in automated testing





Thank you. Questions?

- Mr. Thanasis Kotsiopoulos
- Mr. Alexandros Nizamis
- Emails:
 1. kotsiopoulos@iti.gr
 2. alnizami@iti.gr



CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS