COORDINATEF
Boosting European AI innovation, together.

CitCom.ai label

Alessio Buscemi
Luxembourg Institute of Science and Technology (CitCom.ai)

Funded by
the European Union

A testing use case within Citcom

- The municipal administration of Luxembourg City was experimenting with a third-party provider a citizen-facing chatbot to operate as the first line of interaction between residents and municipal services.

- Before releasing the chatbot, the administration requested LIST to perform an assessment of social biases, focusing on whether the system treats different groups equitably and avoid harmful stereotypes
- a
- The goal was to verify that the chatbot behaves consistently across equivalent prompts, does not make group dependent assumptions, and provides accurate, neutral, and actionable guidance to all citizens.

- Therefore, using our AI Sandbox, we conducted an extensive bias assessment of the chatbot using use case relevant challenges co-designed with the City administration

- We initially identified a few issues, which were solved with the addition of safety guardrails

Good, but...

1. How can the City ensure that the positive results of this one-off bias assessment can be reproduced consistently in future updates or in other municipal chatbots?

2. How can third-party providers and municipalities demonstrate to citizens that such bias testing meets a recognised, minimum standard rather than relying on an ad-hoc evaluation?

3. How can policymakers and administrators communicate transparently that the chatbot has passed a trustworthy and independently verifiable assessment, beyond simply stating that "the issues were fixed"?

The importance of AI labels

AI labels are a way solve the communication gap between technical and non-technical stakeholders [1]:

Make AI understandable

Translate complex technical details into clear, visual summaries for non-experts.

Bridge communication gaps

Help technical and non-technical stakeholders make shared, informed decisions.

Increase transparency and trust

Show key performance, data, and robustness indicators clearly.

Promote responsible use

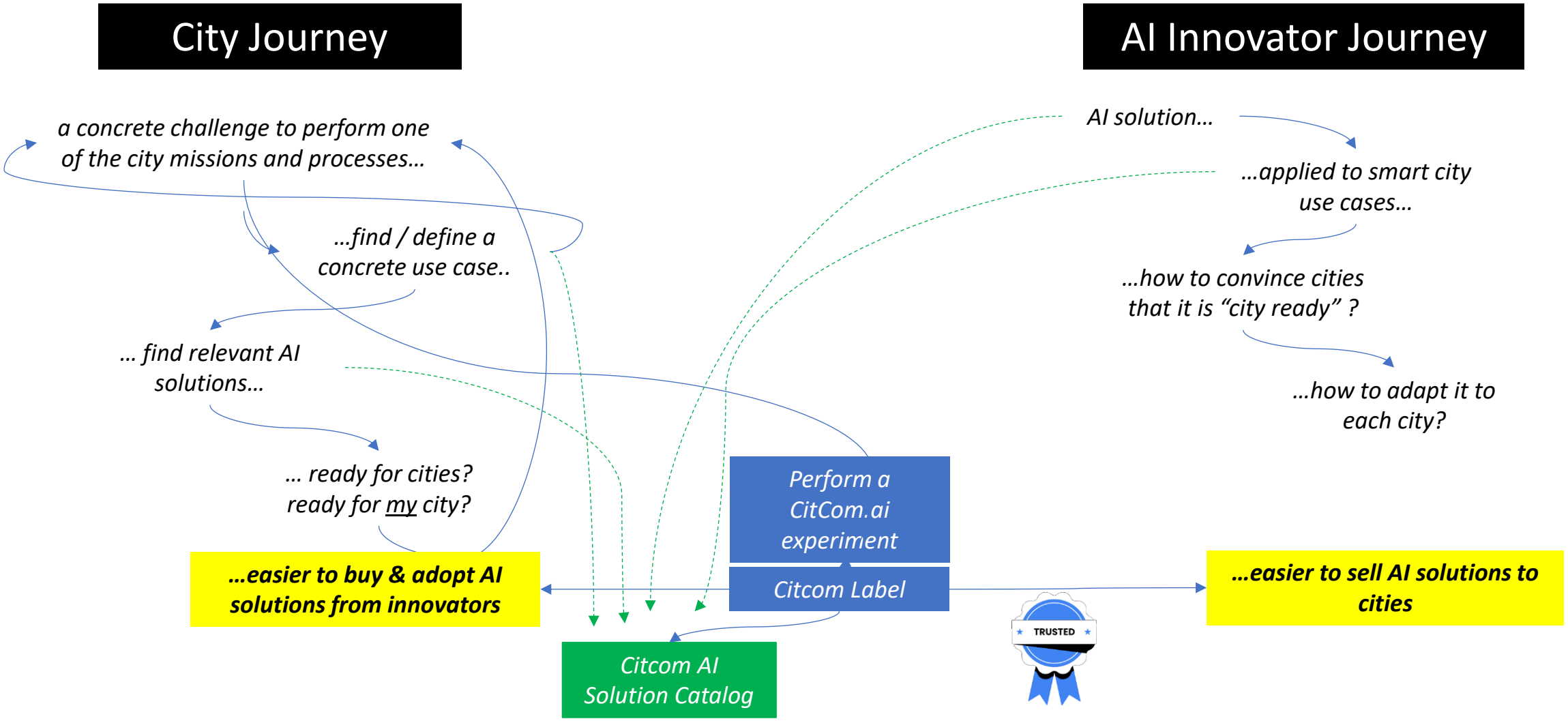
Highlight ethical and sustainability aspects like fairness and energy efficiency.

Support better governance

Offer insights that complement detailed AI documentation.

[1] K. Poitz, R. Fischer, "AI Labels – How They Can Provide Guidance and Build Trust", LAMARR Institute, 2025

CitCom.ai: where cities and AI innovators find each other …… and create mutual trust



A design thinking approach

Name: Thomas De Meester
Organization: imec
Role: Innovation manager

AI and AI assessment-related assets and activities: Belgian node coordinator
Expectations about the AI Assessment club: Looking to contribute imec expertise and use outcomes of assessment club to develop services in the Belgian node.

Name: Ross Linnette
Organization: City of Mechelen
Role: Citcom.ai project coordinator

AI and AI assessment-related assets and activities: TEF site in Citcom, real life testing in the city of Mechelen
Expectations about the AI Assessment club:

Name: Benjamin Vermeulen
Organization: City of Mechelen
Role: CitCom.ai data analist

AI and AI assessment-related assets and activities: Sensor data/platform expert - Sensor solutions and their applications in policy making.
Expectations about the AI Assessment club:

Name: Vicent Botti
Organization: VRAIN - UPV
Role: Full Professor

AI and AI assessment-related assets and activities: member of the AI Safety Institute Consortium (AISIC), member of the project 'Foundations of Trustworthy AI - Reasoning, Learning and Optimization (TAILOR)', expertise in explainable AI and AI Evaluation

Design thinking session (35 Min)

To be compiled offline + 10 min discussion

Have you/your TEF site conducted or required assessments of AI solutions?

YES - Please share a high-level overview, key learnings, open questions, and future directions.

NO - Please explain the reasons, where you believe assessments could be integrated, and your thoughts on future directions.

25 min = 10 min post-its + 15 min discussion

Who are the personas we aim to serve with AI assessments?
What is the value of AI assessment solutions within Citcom?
What kind of AI assessments should be included in the Citcom.ai assessment catalogue?

Prioritization, wrap-up and next steps (15 Min)

- Summary
  - Current state in Citcom
  - Towards an AI assessment catalogue
    - Target personas
    - Value Proposition
    - Type of assessments in the catalogue
    - Catalogue format
- Lead-in to the next session
  - Categorization of different AI assessment approaches.
  - Evaluation of the feasibility and value of a common AI assessment framework.
  - Establishment of asset-sharing mechanisms and defining cross-TEF collaborations for test development.

Expected outcomes

- We have identified key deliverables to guide the next phase of development:
  - **Launch of the AI Assessment catalogue**
  - **Formalisation of Guidelines and Evaluation Reports**
  - **Creation of the Citcom Label**
  - **Pilot Implementation**

Launch of the AI assessment catalogue

- A fundamental step before the actual creation of the catalogue, is the identification of categorisation guidelines for testing solutions, in a way that allows easy matching with the use cases requiring the assessment

- This are the criteria we identified:

Current assessment capabilities across Citcom

Solution name	Provider	Licensing Type	Project Phase/ TRL	Domain of Application	Ethical Dimensions	Security & Securitization of Data	Assessment Type	Example of use case	Resources

Citcom experiments requiring AI Assessment

Experiment	City	Description	AI Innovator	Project Phase/ TRL	Assessment requirements	AI Risk Category	Ethical Dimensions	Security & Securitization of Data	Assessment Type	Resources

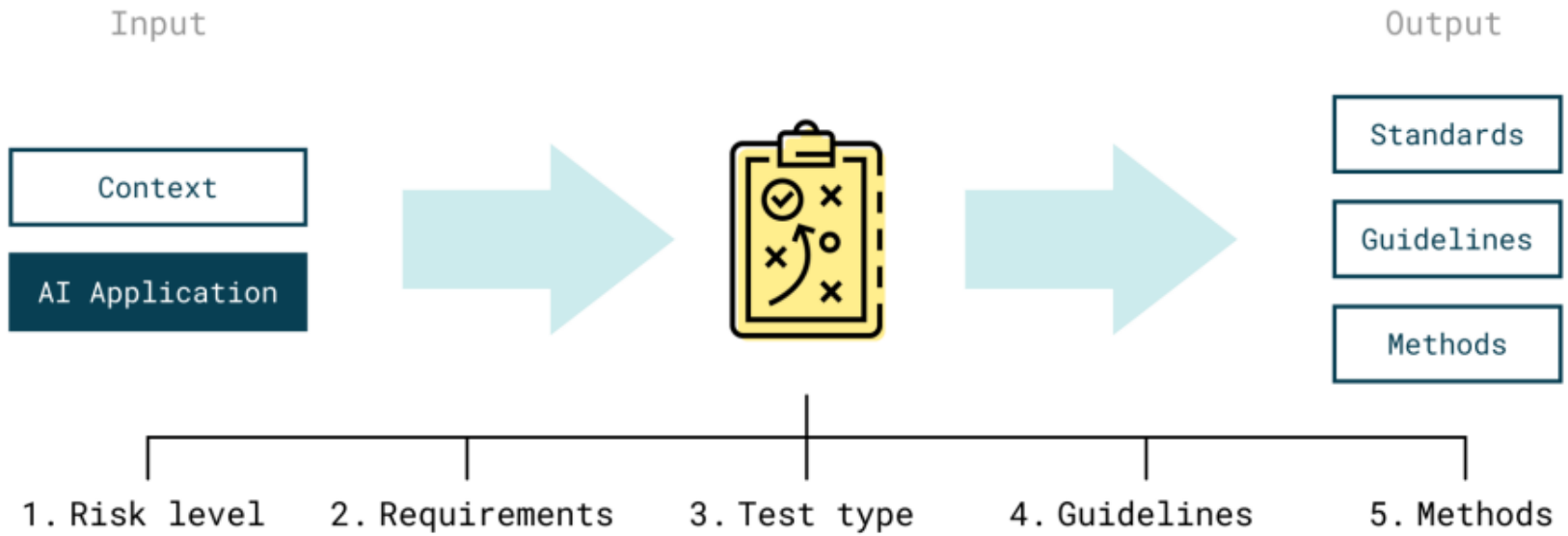
- The next step is to create the actual catalogue collecting all assessment solutions operated by Citcom partners

Expected outcomes

- We have identified key deliverables to guide the next phase of development:
  - **Launch of the AI Assessment catalogue**
  - **Formalisation of Guidelines and Evaluation Reports**
  - **Creation of the Citcom Label**
  - **Pilot Implementation**

Methodology

- LIST and RISE started the mapping from legal requirements to methods, based on RISE's methodology*



- In our soon-to-be-released paper, we work on identifying:

Requirements | Sources | Assessment Dimensions

*Mowla, N. et al., (2024). *From AI Act to structured testing of AI systems*. RISE Research Institutes of Sweden.

Requirements

- We identified 11 categories of requirements that capture the key dimensions of AI trustworthiness & compliance

- The starting point was the 7 principles of Trustworthy AI defined by the European Commission's HLEG:

🔎 **Human agency and oversight**

🛡️ **Robustness and safety**

🔒 **Privacy and data governance**

🔳 **Transparency**

✅ **Accountability**

👐 **Fairness, diversity & non-discrimination**

🌱 **Societal and environmental well-being**

- To complement these ethical foundations with more operational and procedural dimensions, we added 4 additional categories corresponding to key obligations introduced by the AI Act:

Quality management

Risk Management

Technical Documentation

Record-keeping

Sources for controls and testing methods

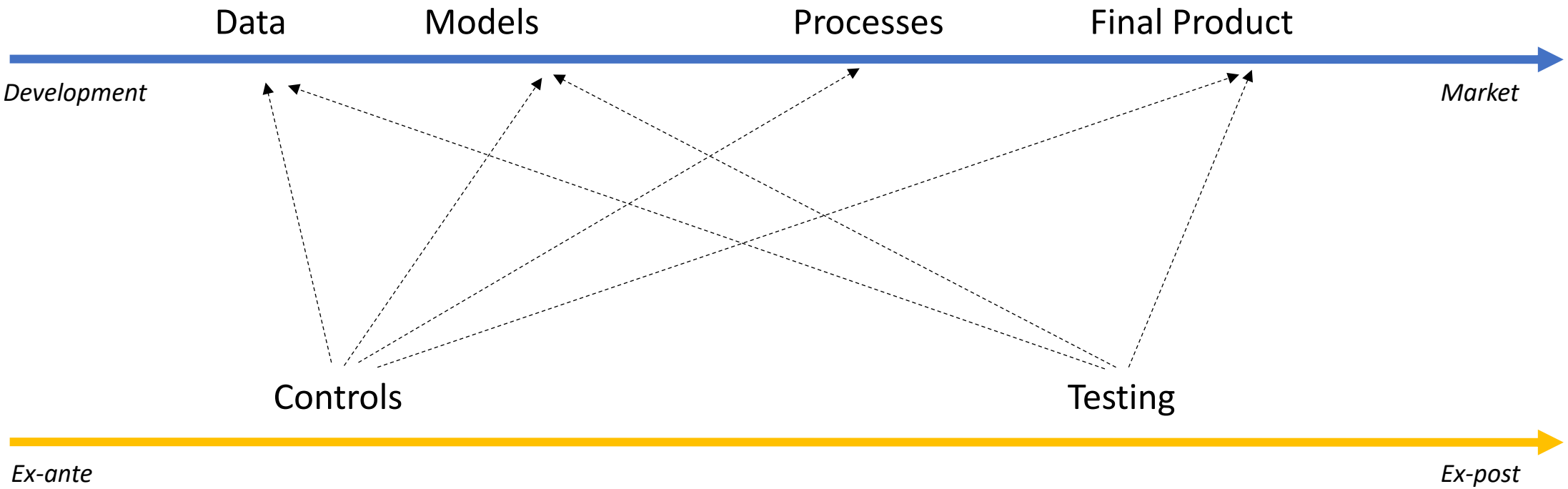
Binding or quasi binding regulatory sources (AI Act, GDPR etc.)

International standards or authoritative guidance (ISO, GPAI Code of Practice, ISACA etc.)

Recognised and highly cited scientific work

Dimensions of Assessment

AI Traceability



Type of assessment

Expected outcomes

- We have identified key deliverables to guide the next phase of development:
  - **Launch of the AI Assessment catalogue**
  - **Formalisation of Guidelines and Evaluation Reports**
  - **Creation of the Citcom Label**
  - **Pilot Implementation**

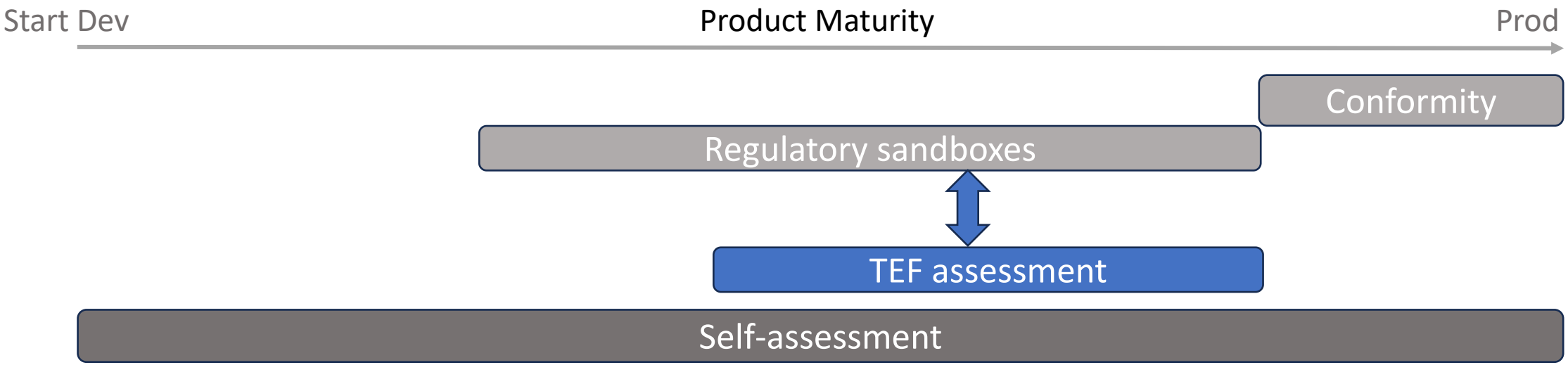
Proposed Citcom Label

- **Value proposition:** Independent third-party assessment of AI trustworthiness, with **non-binding compliance recommendations** to guide cities and procurement officers.
- **Badging system:** Result-specific badges awarded to innovators, with granularity based on market needs.
- **Credibility & thresholds:** Consistent credibility across badges ensured by harmonised guidelines andcase-by-case expert evaluation.
- **Badge infrastructure:** Timestamped, tamper-proof badges linked to the Citcom Hub for verification,transparency, and metadata access.
- **Reporting & visibility:** Harmonised evaluation reports with **legal disclaimers**; public list of participating AI innovators hosted on the Citcom Hub.

Expected outcomes

- We have identified key deliverables to guide the next phase of development:
  - **Launch of the AI Assessment catalogue**
  - **Formalisation of Guidelines and Evaluation Reports**
  - **Creation of the Citcom Label**
  - **Pilot Implementation**

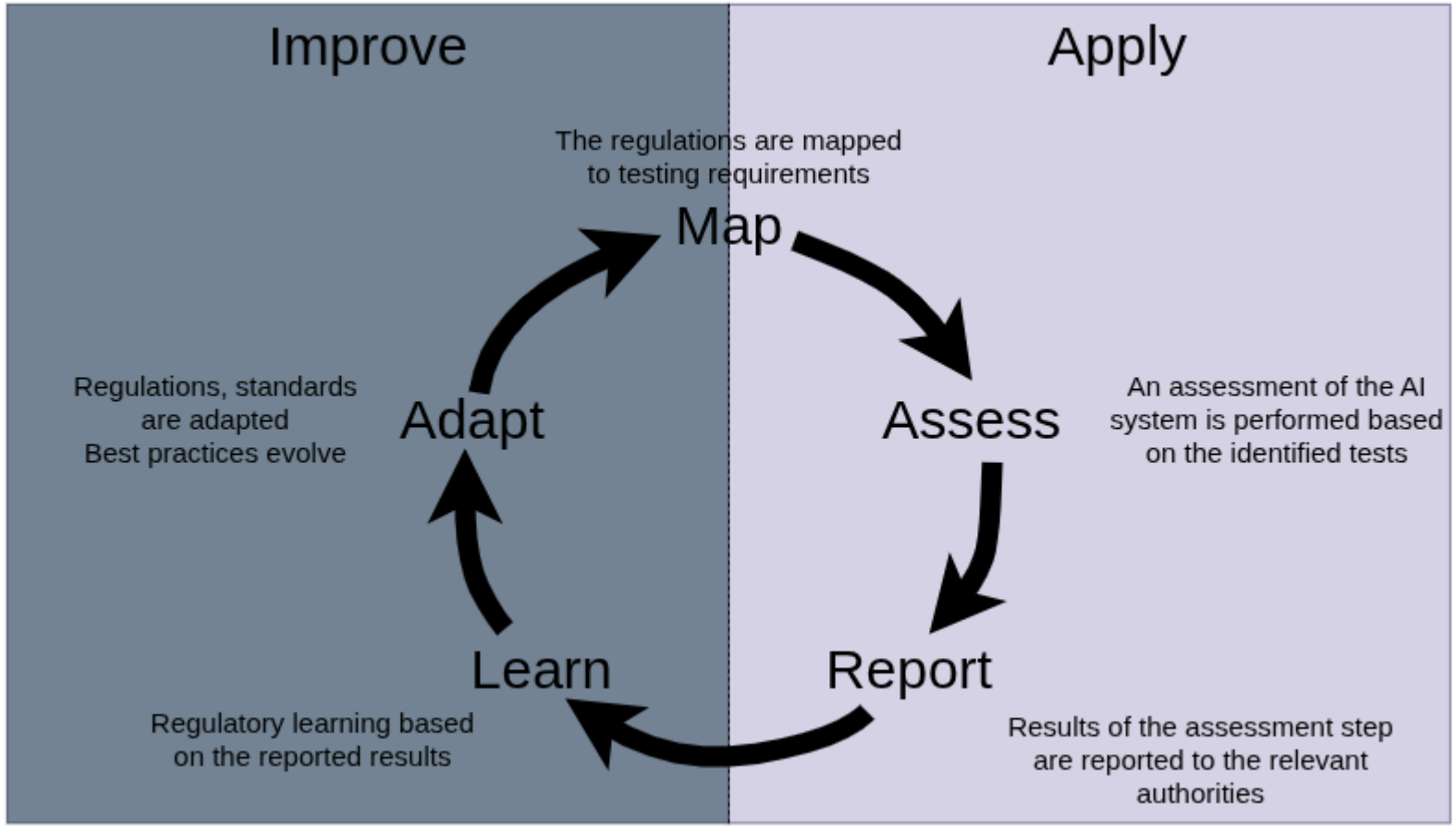
TEF assessment in the broader EU innovation ecosystem



- We believe that TEF assessment could be also valuable with respect to AI Regulatory Sandboxes (AIRSes)
- The AI Act (Art. 58(2)(i)) explicitly mandates that AIRSes should *"facilitate the development of tools and infrastructure for testing, benchmarking, assessing and explaining dimensions of AI systems [..]"*
- Testing tools, thresholds, and result interpretations must be adapted to the specific sector and use case
- TEFs could serve as preferred partners for Competent Authorities seeking external technical expertise to carry out assessments in Regulatory Sandboxes within their respective domains

MARLA

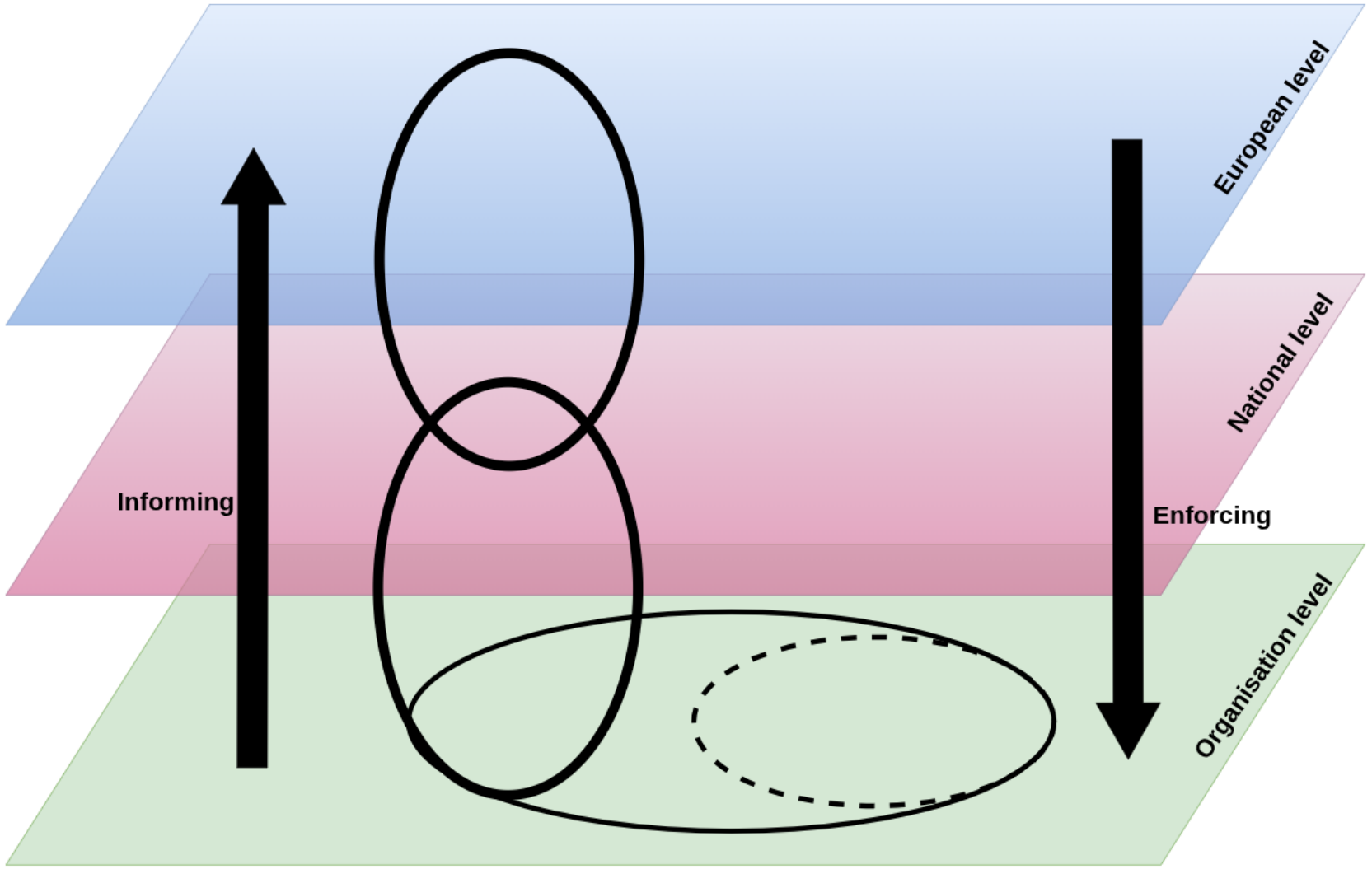
- Drawing from discussions and personal experience with regulators and legal experts, fostering effective communication between technical and legal professionals remains a major challenge.
- This dialogue is essential for practitioners to correctly interpret legal requirements and to foster regulatory learning, a key principle introduced by the AI Act.
- For this reason, within the Luxembourg AI Factory, we developed **MARLA**, a high-level framework designed to identify and structure the various stages of this cycle.



Enabling dialogue between developers and regulators

MARLA can be implemented on three complementary levels:

- **Organisational level:** to support internal learning and continuous improvement within companies or institutions.
- **National level:** to enable regulatory learning and structured dialogue with competent authorities.
- **European level:** to promote alignment, knowledge exchange, and coherence across Member States.



Conclusion

- The Citcom label is conceived as an independent third-party evaluation of AI trustworthiness, offering non-binding compliance recommendations to guide cities and procurement officers in their decision-making.
- It can serve as a blueprint for other TEFs
- The assessments conducted within the TEFs can serve as a benchmark in the European AI landscape for the sectors they cover.
- To ensure coherence and complementarity across domains, other TEFs are encouraged to contribute to the refinement of the assessment guidelines.
- Where sector-specific regulations apply, their additional requirements can be mapped and integrated into the proposed methodology.