.DATAWEEK24
JOIN.LEARN.SHARE.GET VALUE

Data Spaces Symposium

14:00

AI Testing and Data Spaces

Petra Dalunde, Kateryna Mischenko,
Nishat Mowla, Giovanni Leoni, Daniel Sáez-Domingo & Yunus Bulut

DATA SPACE SYMPOSIUM DARMSTADT 2024

AI Testing & Data spaces

Petra Dalunde – Coordinator AI Testing at RISE

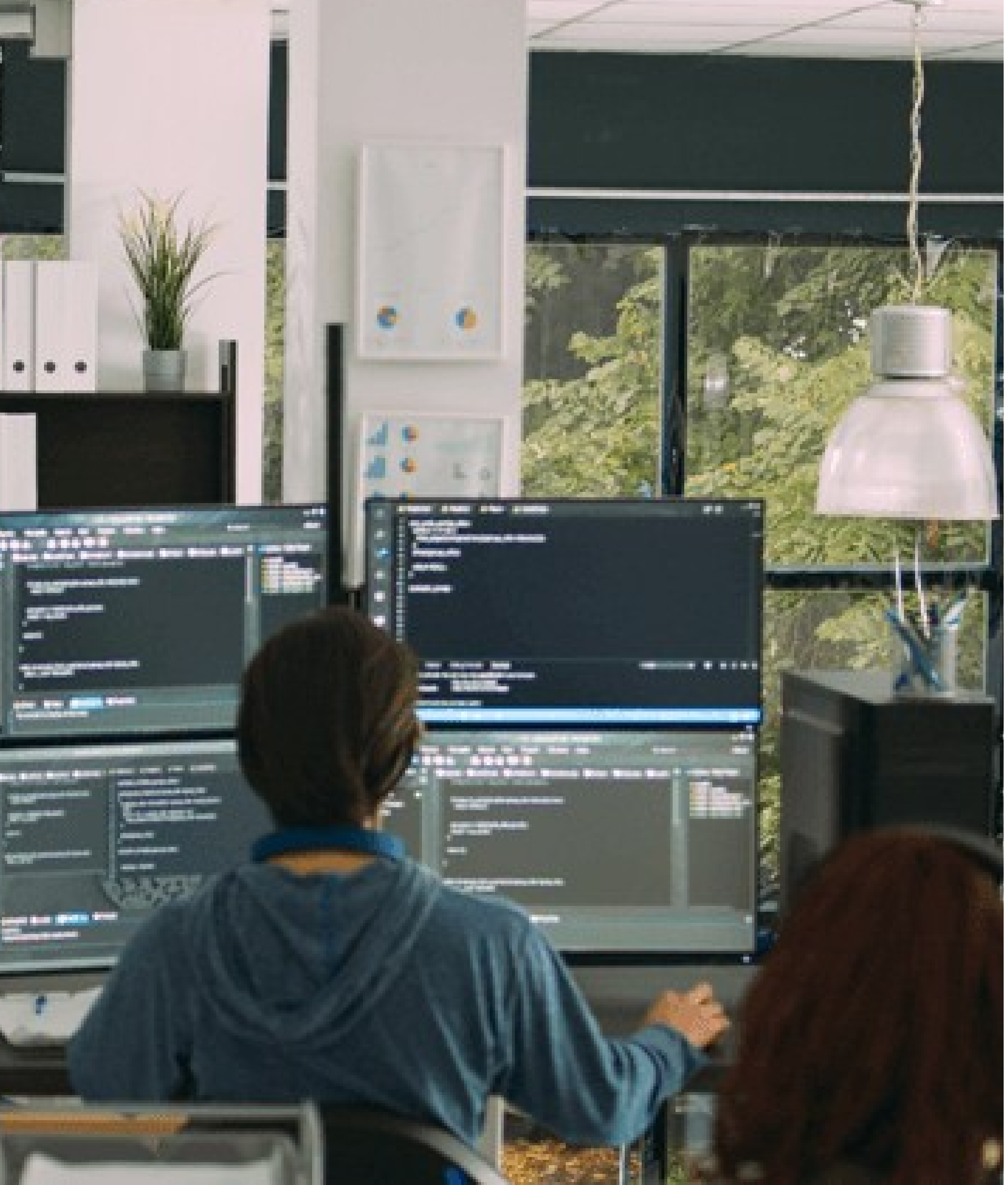
RISE RESEARCH INSTITUTES OF SWEDEN

- RTO
- 3 TEFs
- Test & Demo
- Notified body
- i-space in BDVA

EDIHs – TEFs – Cyber Security – Industry - Medicine

AI Testing in this session

”Testing AI systems towards legislation and function for human centric AI in Europe”



Todays session

Exploring AI Testing: Introduction and Methodology
Katya Mishchenko & Nishat Mowlat at RISE

AI Governance in organizations: Data, Process & People to make it happen in practice
Giovanni Leoni, at Credo AI

i-Spaces and Sandbox
Daniel Sáez Domingo at ITI

AI testing in practice: What Validator has learnt from its customers
Yunus Emrah Bulut at Validator

Panel & Q&A

Exploring AI Testing: Introduction and Methodology

March 12th, 2024

Kateryna Mishchenko & Nishat I Mowla

RISE

Agenda:

- Introduction to AI testing:
  - What, why and how?
  - About AI Act and standards
- Challenges in AI testing
- AI testing methodology
  - Assessment list of Trustworthy AI
  - AI Testing standards
  - Application domains and subfields of AI
- Performing AI testing
  - AI testing at RISE

Introduction to AI testing

About AI testing

What:

Testing AI systems is s a vital part of the development and deployment of AI systems since ensures their accuracy, reliability, safety, efficiency and effectiveness

Why:

AI testing builds trust and confidence in real-world applications and helps in identifying and rectifying potential issues early, thereby improving the quality of software releases.

How:

One instrument to emphasize the importance and ensure the safety and reliability of AI systems is the **AI Act** (enters into force 2024-2026).

It lays down harmonized rules on AI, aiming to balance the socio-economic benefits and potential risks of AI technologies placed on the European market.

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

AI Act: risk-based approach

UNACCEPTABLE RISK:
AI systems prohibited, because considered a clear threat to the safety, livelihoods and rights of people

HIGH-RISK:
AI system subject to strict obligations and a conformity assessment before they could be put on the market

LIMITED-RISK:
AI system with specific transparency obligations

LOWER-RISK:
They represent the vast majority of AI system with minimal risk or no risk, allowed to be used freely without any obligation from the discussed regulation

Source: https://www.iasonltd.com/doc/jit/2021/European_Commission_Regulation_on_AI.pdf

About the Standards related to AI Act

- **ISO/IEC TR 29119-11:2020 Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems**

  Provides an introduction to AI-based systems, new challenges and opportunities for testing them.

  This document explains those characteristics which are specific to AI-based systems and explains the corresponding difficulties of specifying the acceptance criteria for such systems.

- **ISO/IEC AWI TS 29119-11 Software and systems engineering — Software testing — Part 11: Testing of AI systems**

  Describes testing techniques applicable for AI systems in the context of the AI system life cycle model stages

  Shows how AI and ML assessment metrics can be used in the context of those testing techniques. It also maps testing processes to the verification and validation stages in the AI system life cycle.

- **ISO/IEC 25059 Software engineering. Systems and software Quality Requirements and Evaluation (SQua**RE)

  Outlines a quality model for AI systems and provide guidelines for measuring and evaluating the quality of AI systems, focusing on characteristics like accuracy, interpretability, robustness, fairness, privacy, and security.

RI. SE

Challenges in AI Testing

Some challenges related to AI testing

- Testing AI systems comes with unique challenges, such as the unpredictability of AI behaviour, the difficulty in defining the right metrics for success, and the complexity of creating diverse and representative test cases.
- ISO/IEC AWI TS 29119-11 "Software and systems engineering — Software testing — Part 11: Testing of AI systems" describes testing techniques and metrics for AI systems in the context of the AI system life cycle model stages. According to it, some of **challenges** are:

Data testing:

issues with data quality, diversity,
privacy, labeling,
temporal sequencing,
data drift, and potential
biases.

Explainability:

Arises from “black box”, nature, making it difficult to understand why they make certain decisions.

Continuous Learning:

often learn and adapt over time, which means they need to be continuously tested and monitored

Transparency:

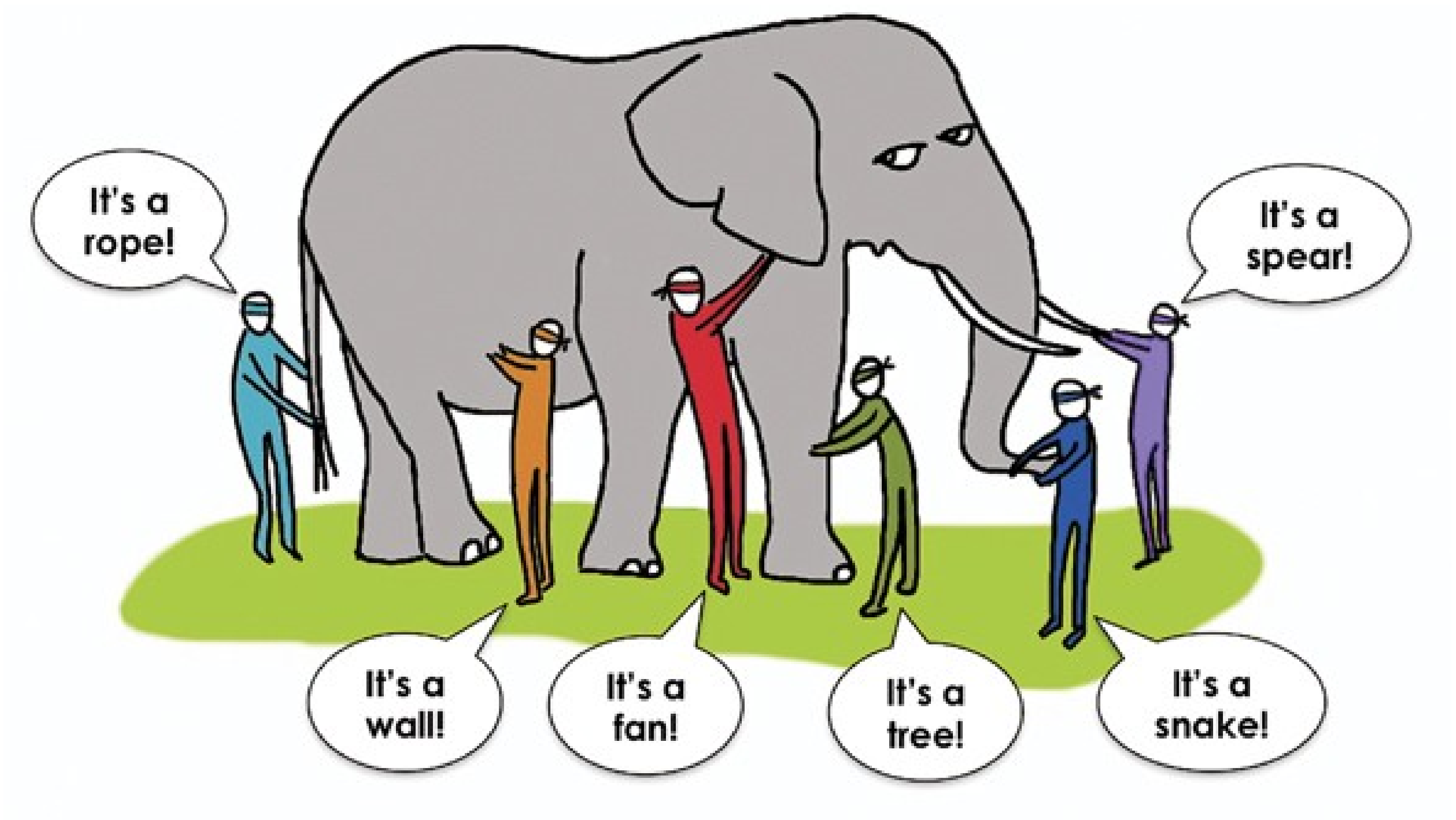
arises “black box” nature, sensitivity of training data, dynamic learning, potential for bias, and the trade-off between model accuracy and explainability.

Trustworthiness:

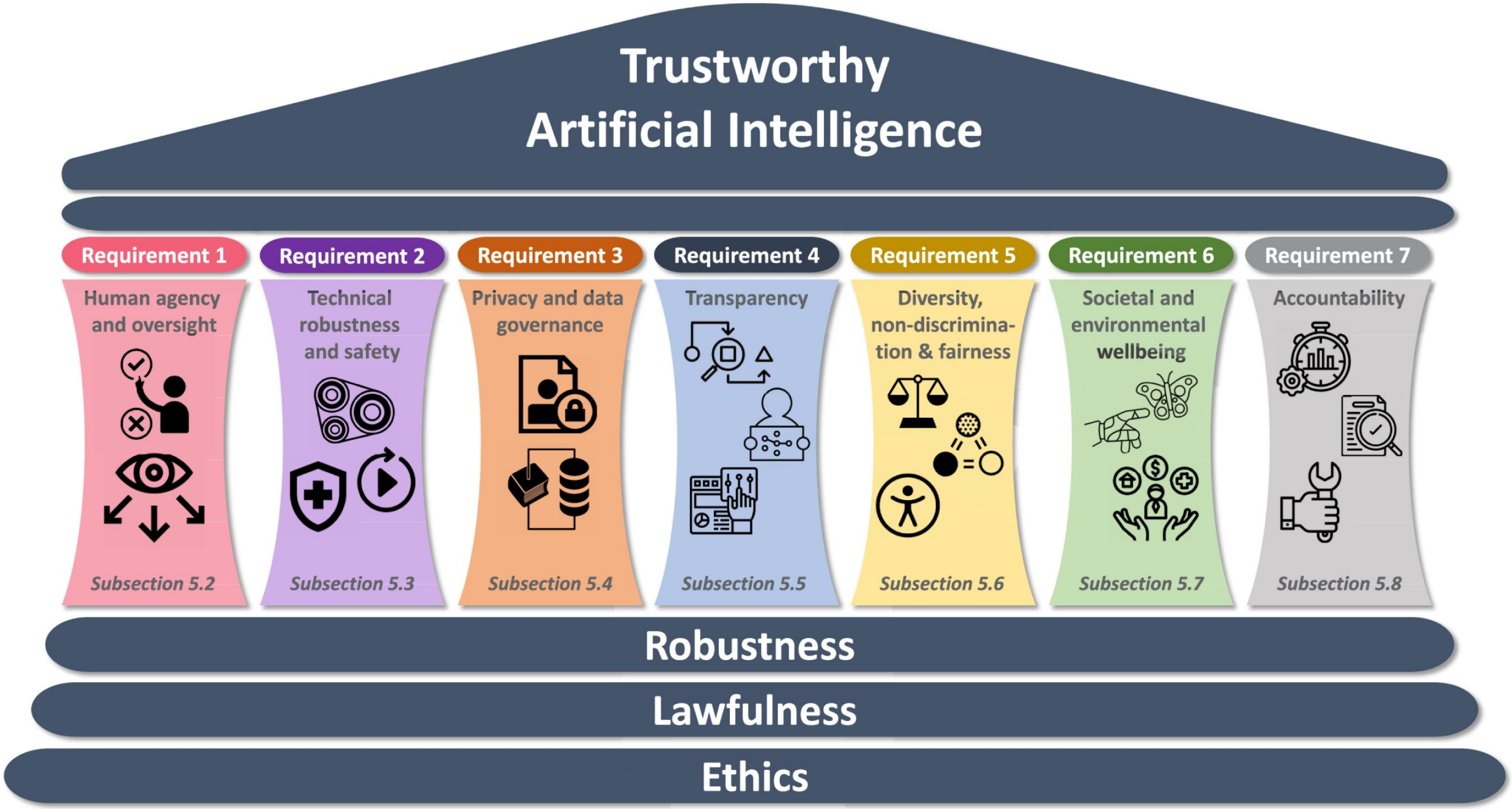
arises from the “black box” nature, the need for security against manipulation, the requirement for data privacy, the necessity for accountability, and the complexity of ensuring fairness and non-discrimination.

AI testing methodology

The AI Testing Elephant



Assessment list of Trustworthy AI (ALTAI)



Ethics guidelines for Trustworthy AI: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Image source: https://www.sciencedirect.com/science/article/pii/S1566253523002129

AI Testing Standards

Classification and evaluation	AI Software quality	Security, trustworthiness, privacy	Safety	Data quality & bias	Robustness and reliability	Ethical and societal concerns	Management & Lifecycle	Risk management
ISO/IEC 29119 series	ISO/IEC 23053	ISO/IEC 22989	ISO/IEC 22989	ISO/IEC 5259	ISO/IEC 27001	ISO/IEC 24368	ISO/IEC 42001	ISO/IEC 23894
ISO/IEC 4213	ISO/IEC 24028	ISO/IEC 20547	ISO/IEC 5469	ISO/IEC 24027	ISO/IEC 24029		ISO/IEC 42006	ISO/IEC 31000
ISO/IEC 25059	ISO/IEC 25000	ISO/IEC 24028					ISO/IEC 38507	
ISO/IEC 42102							ISO/IEC 5338	
Functional		Non-functional						

Application domains and subfields of AI

Subfields of AI:

1. Machine learning
2. Deep learning (DNN)
3. Natural language processing (LLM)
4. Computer vision (image, video, voice)
5. Reinforcement learning (agents)
6. Multi-agent systems
7. Robotics (autonomous)
8. Expert systems (reasoning)
9. Speech processing (speech recognition)
10. Planning and scheduling (plan actions)
11. Knowledge representation and reasoning
12. Evolutionary computing (genetic algorithm)
13. Affective computing (recognize feelings)

Performing AI Testing

Performing AI Explainability Testing

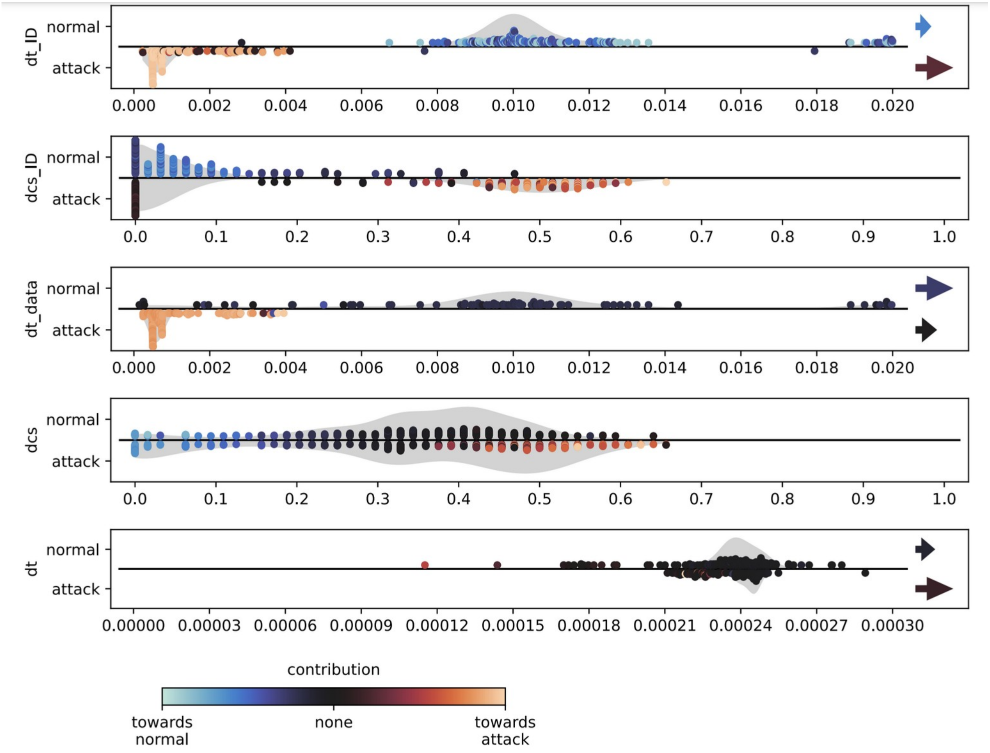


FIGURE 2. VisExp | A pseudo-global visualization-based explanation, using SHAP values. It shows the features in the dataset in swarm plot-like strips for normal and attack classifications. Each point is an instance from the train data. The x-axes are the feature values, and the color represents the SHAP values. The color of the arrows represent the mean of the SHAP values outside of the diagram, and their relative size represents how many data points there are.

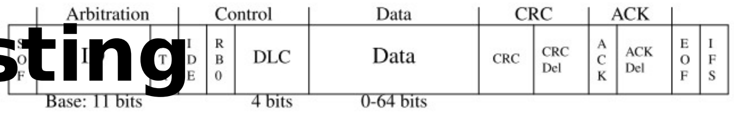


FIGURE 1. CAN frame | The Survival dataset has features of the ID, DLC and data field, along with the timestamp of when a CAN frame is transmitted.

TABLE 1. DNN hyperparameters | Parameters and their values as specified when building the DNN in keras.

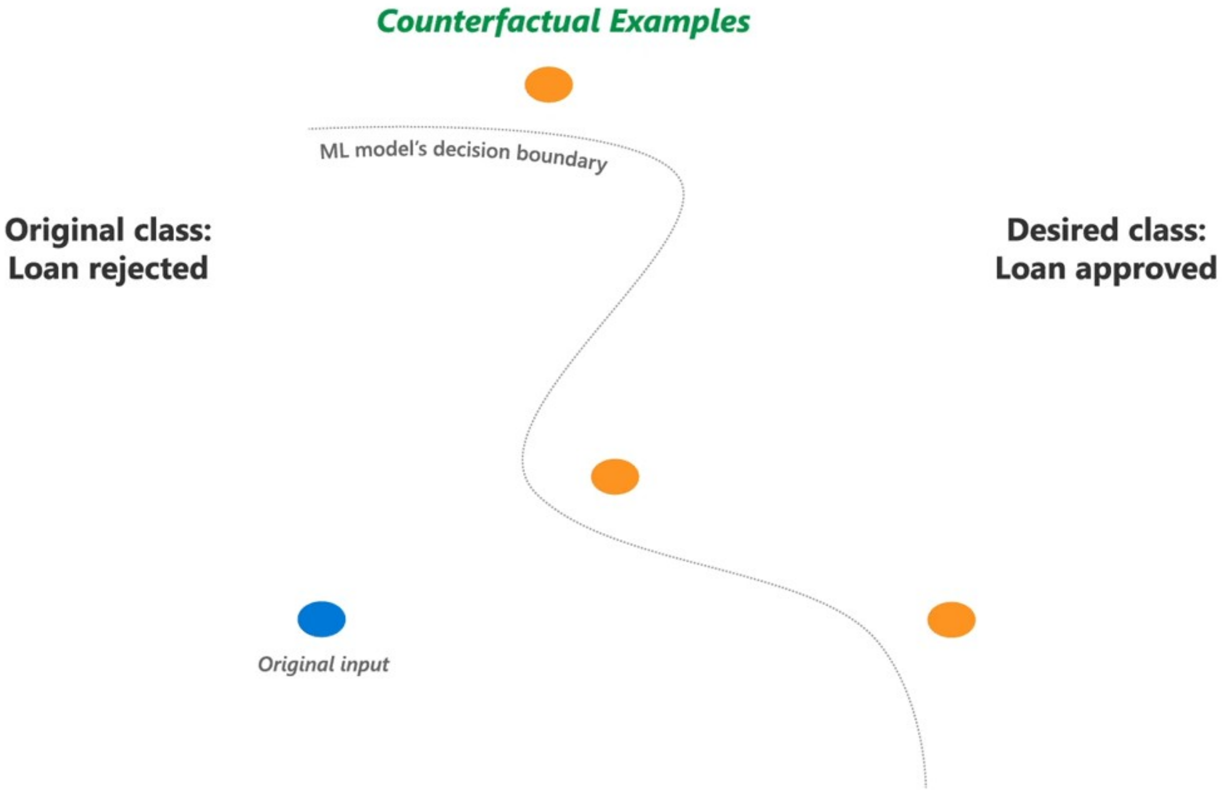
Layer	# of units	Description
`layer_1`	11	`keras.layers.Dense`
`layer_2`	23	`keras.layers.Dense`
`layer_3`	7	`keras.layers.Dense`
Hyperparameter	**Value**	
`optimizer`	"adam"	Optimizer algorithm
`batch_size`	200	# of samples in a gradient descent
`epochs`	20	# of training passes over the dataset

TABLE 2. The engineered features.

Feature	Description
dt [12]	Transmission time (s) between CAN frames
dt_ID [12]	Transmission time (s) between CAN frames with the same ID
dt_data	Transmission time (s) between CAN frames with the same data field
dcs	Data change score (ratio) between CAN frames
dcs_ID	Data change score (ratio) between CAN frames with the same ID

Hampus Lundberg, Nishat I Mowla, Sarder Fakhrul Abedin, Kyi Thar, Aamir Mahmood, Mikael Gidlund, Shahid Raza, "Experimental Analysis of Trustworthy In-Vehicle Intrusion Detection System Using eXplainable Artificial Intelligence (XAI)," IEEE Access, vol. 10, September 2022. (Link)

Performing AI Explainability Testing



[Microsoft Research]

Quality of AI

Quality AI requires quality data

But quality AI is more than data

- **Cybersecurity**
- **Transparency**
- **Robustness**
- **more**


Thanks!

kateryna.mishchenko@ri.se
nishat.mowla@ri.se

GIOVANNI LEONI

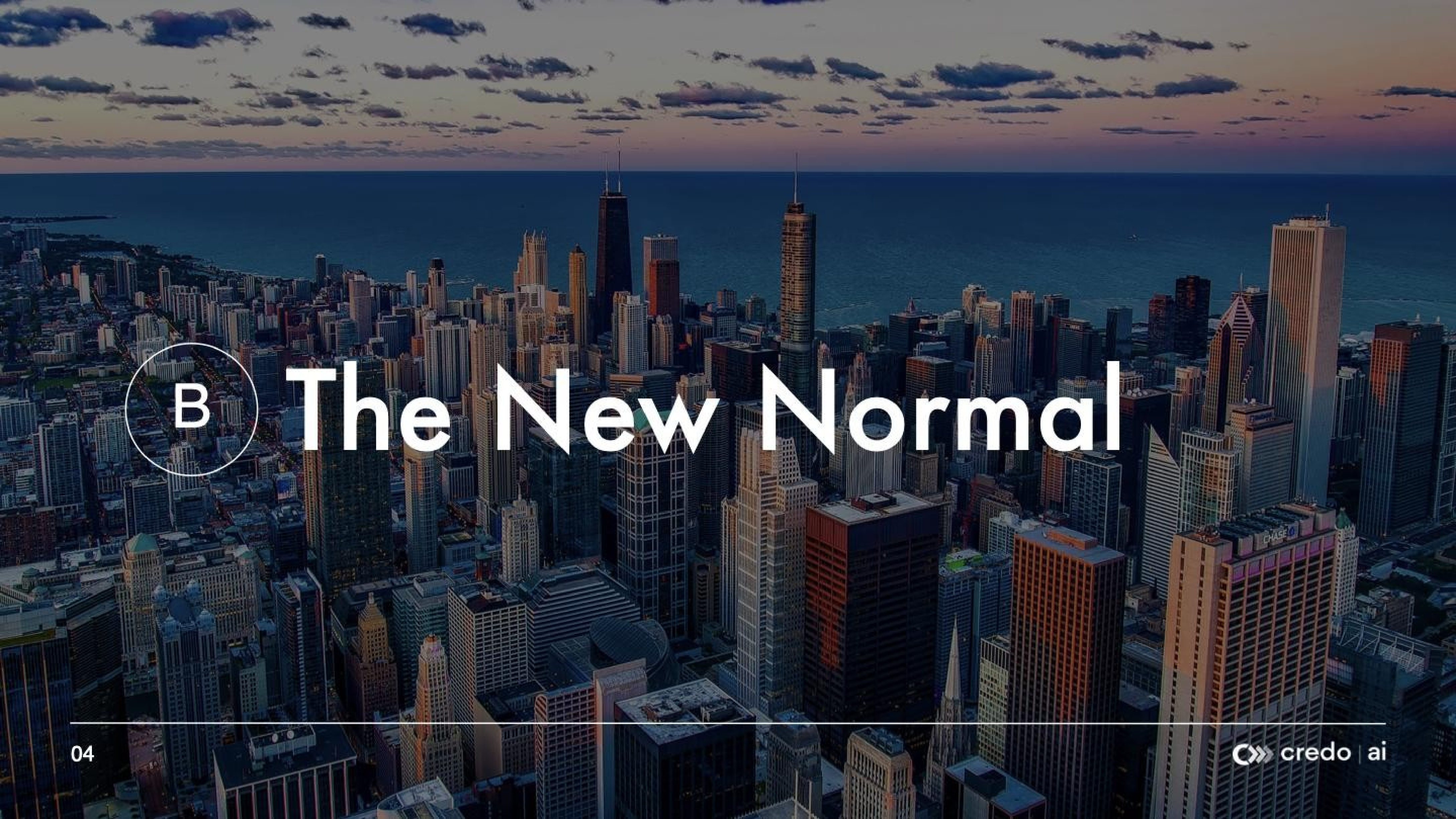
AI Governance in organisations

Data, Process & People to make it happen in practice

Agenda

- A — AI Testing & Data Spaces
- B — The New Normal
- C — The Business Case
- D — Building Capability
- E — The Future of Trustworthy AI

Data is the fuel and Data Spaces creates the preconditions for innovating and adopting Trustworthy AI



B The New Normal

A The New Normal

Expectations are on the rise.

A The New Normal

EU AI Act is not only a regulation.

C The Business Case

B

The Business Case

lack of control

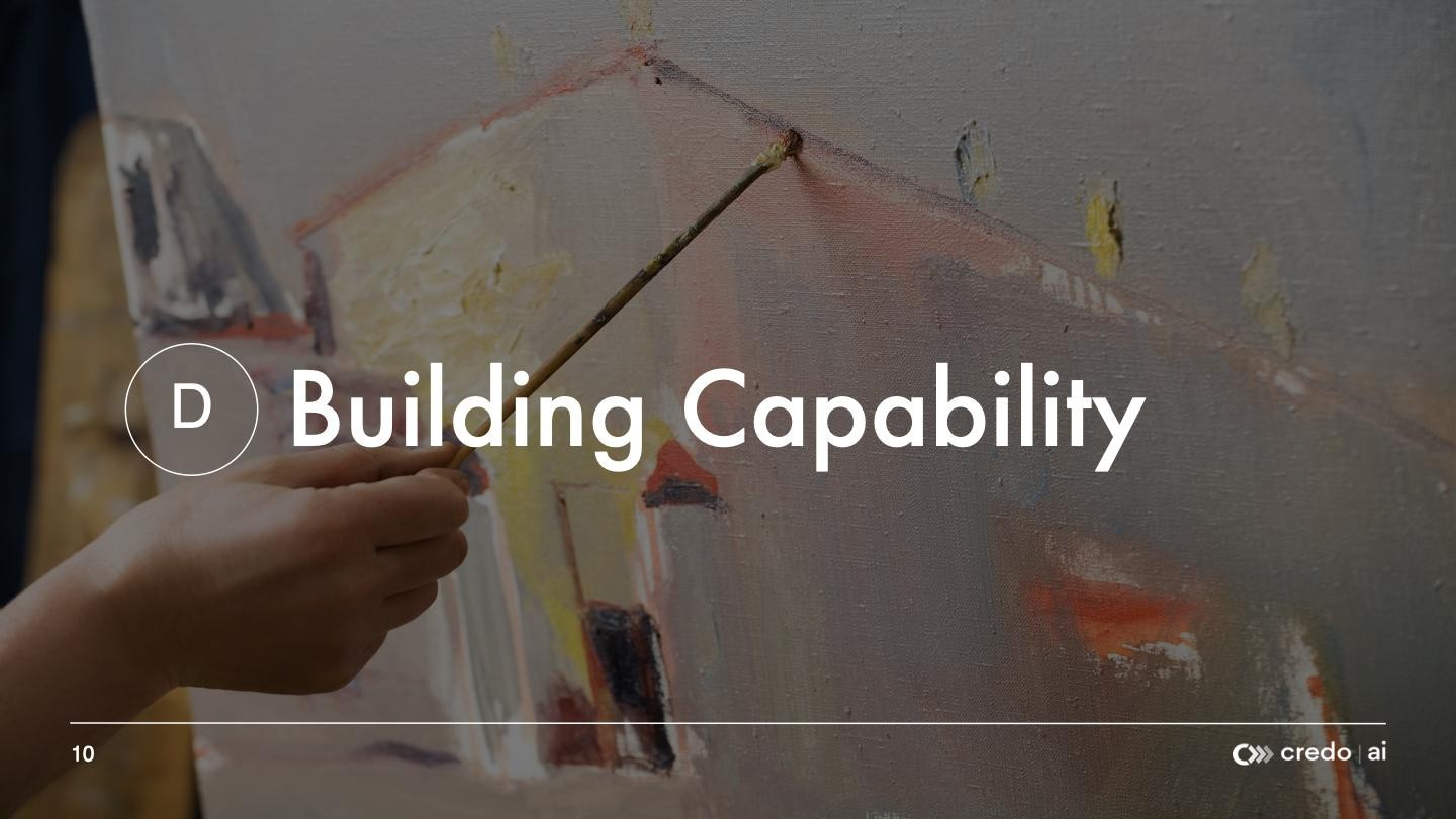
inefficiencies

brand risk

liabilities

B The Business Case

confidence in AI	informed accountability
adapting to the world	trust in digital



D Building Capability


C
Building Capability
Strategy & Policy
Blueprint
Engagement
Building through doing
credo ai

E

The Future of Trustworthy AI enabled by Data Spaces

It's not about perfection; it's about starting, today.

GIOVANNI@CREDO.AI

CREDO.AI

i-SPACES AND SANDBOXES

ITI EXPERIENCE

ITI INVESTIGATE TO INNOVATE

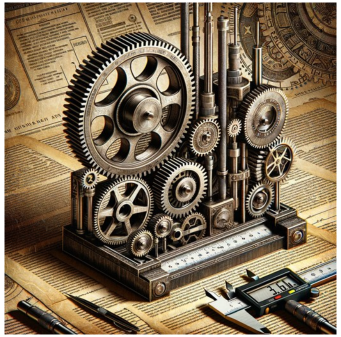
Daniel Sáez-Domingo :

(dsaez@iti.es)

- Strategic Intelligence and Technology Transfer Director in **ITI**
- Coordinator of ITI Innovation Space (**i-Space Platinum BDVA**)
- Member of Board of Directors of **BDVA and GAIA-X**

Data Driven economy needs

Tech. Convergence



Regulations & Standards

Awareness

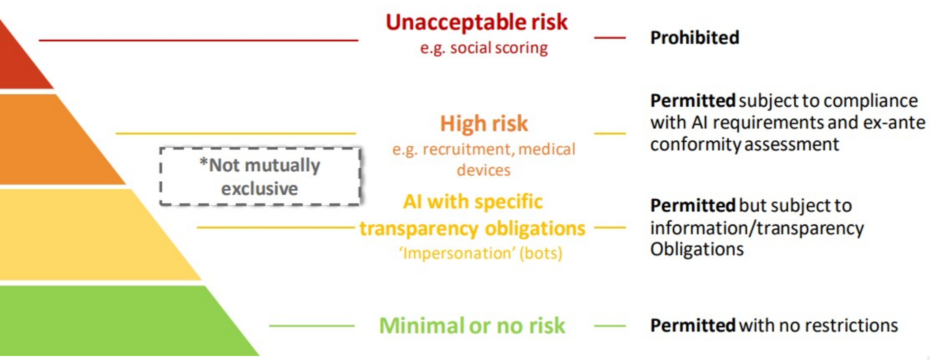
Experimentation

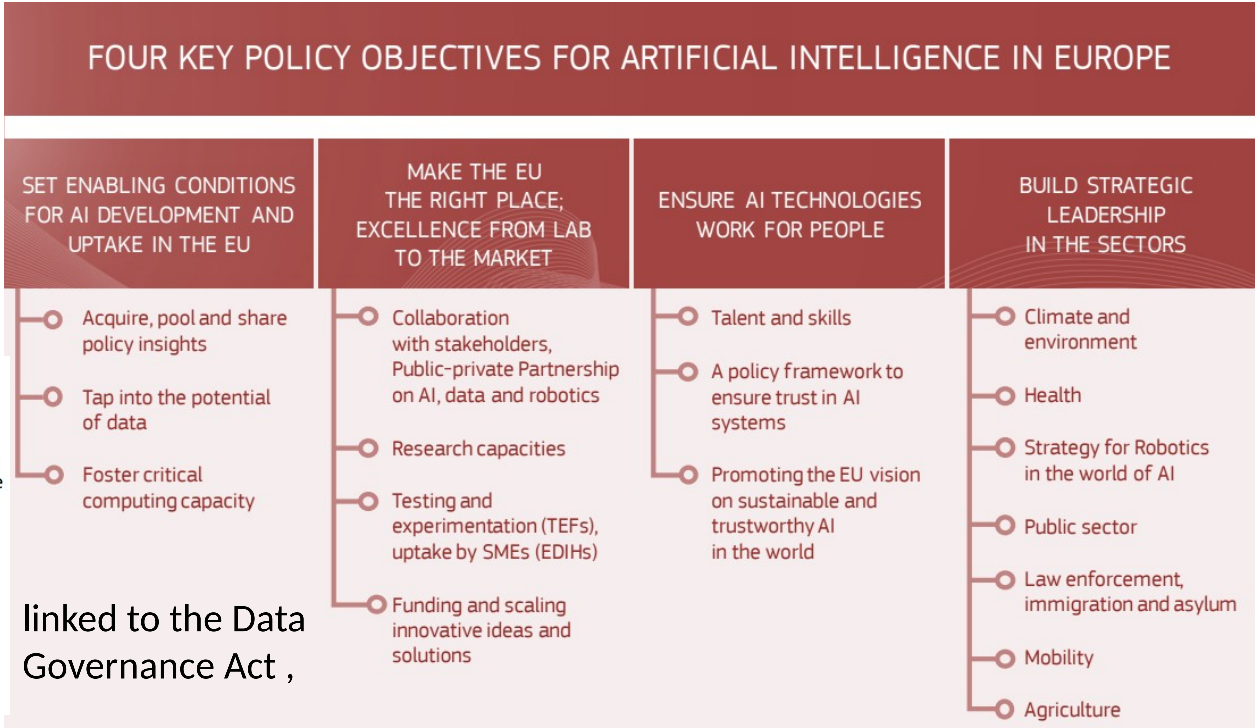
i-Spaces, making Data Strategy happen

i-Spaces concept was defined a long time ago (2014), but it is totally alive and needed nowadays. i-Spaces are ecosystems with **powerful infrastructures, knowledge, tools, data**, … ready to provide services for the **experimentation and innovation with Data and AI.**

They have a close contact and are well known and recognized in their local ecosystems and are also very well connected globally, as stars of an impressive constellation around Data and AI.

EU AI Act





linked to the Data Governance Act ,

The New Legislative Framework (NLF) legislation (e.g. machinery, medical devices, toys), will integrate a conformity assessment procedures for AI risk classification

Sandboxes

A Sandbox is an isolated framework to allow innovators, whether start-ups or large firms, to conduct live experiments in a controlled environment

Tech. Challenges

Legal Challenges

Business Challenges

Ethics and Regulatory Challenges

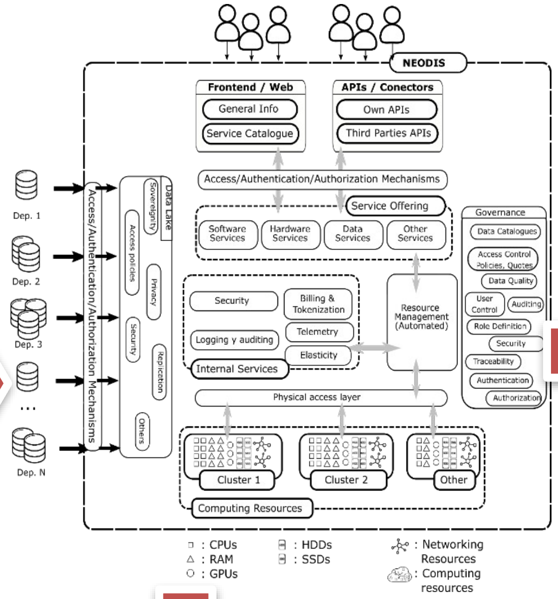
WHY SANDBOXES: EXAMPLE SPAIN

i-Spaces: Trusted environments to experiment

Knowledge for on-boarding in the digitization and extraction of value from data

Model, tools and knowledge to deploy data & Ai stacks

Own infrastructure and tools to host experiments

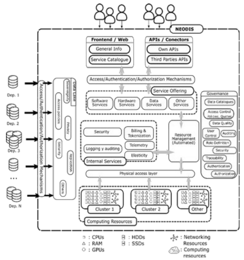


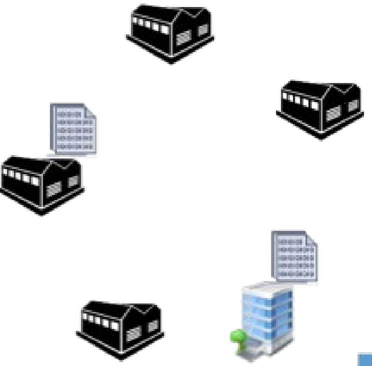
- **Enriched node** providing:
  - Data Quality
  - Data Governance
  - Data Ingestión
  - Data Storage
  - Data Sharing
  - Data Processing
  - Access control and security
- A **Data Innovation Lab**

Access to personalized resources	Access to services	Skills

•Infrastructures
•Datasets
•Tools
	•Data Processing
	•Data Management
	•Business
	•Legal
	•Ethics
		•Technical
		  •Infrastructure deployment
		  •Data governance
		•Business

How can i-Spaces help the regulators





BDV i-Space

Access to personalized resources	Access to services	Skills

• Infrastructures
• Datasets
• Tools
	• Data Processing
	• Data Management
	• Business
	• Legal
	• Ethics
		• Technical
		  ▪ Infrastructure deployment
		  ▪ Data governance
		• Business

SCENARIO 1
- Infra
- Services
- Business

SCENARIO 2
- Infra
- Services
- Business

SCENARIO 3
- Infra
- Services
- Business

AGILE AND PERSONALIZED SCENARIO CONFIGURATION ON DEMAND

How can i-Spaces help the regulators

Analyse the AI system, its intended use, potential risks, and mitigation strategies.

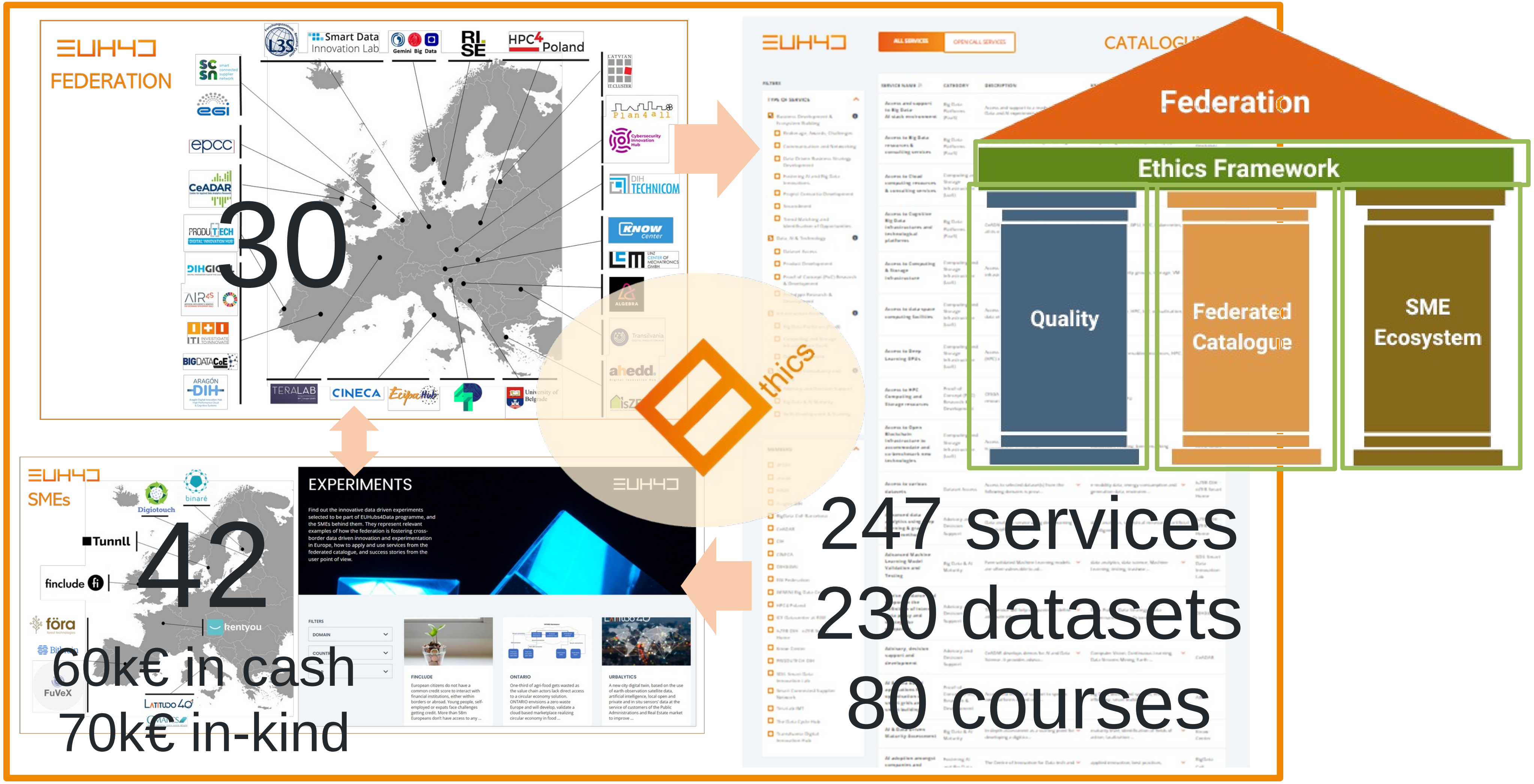
Prepare the scenario to test and deploy the AI systems in real-world settings, but in controlled environment. Limited time.

Report to the regulator on the performance, risks, and any incidents related to the AI systems

And i-Spaces are sharing best practices

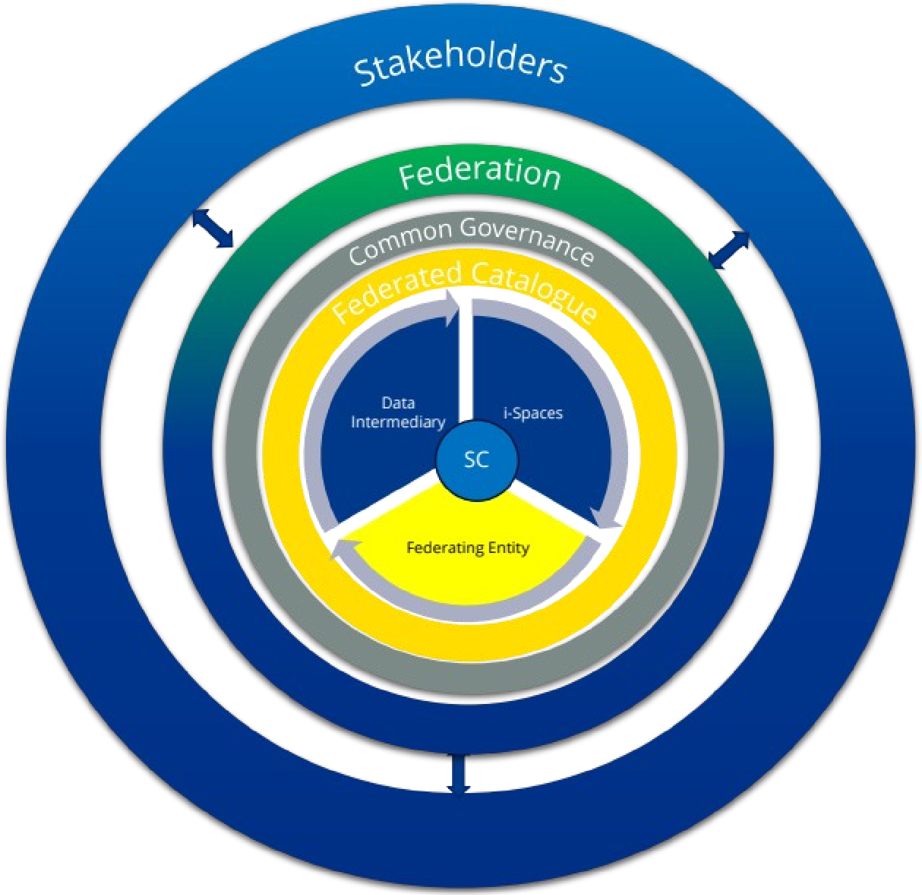
i-SPACES HAVE FEDERATED THEMSELVES WITH THE VISION OF BEING THE REFERENCE FOR EXPERIMENTATION AND INNOVATION WITH INDUSTRIAL, PUBLIC AND PERSONAL DATA AND AI TECHNOLOGIES FOLLOWING THE EUROPEAN, NATIONAL AND REGIONAL VALUES AND PRINCIPLES. SOLID COLLABORATION. COMMON VISION. TANGIBLE IMPLEMENTATION

4 main pillars



Mission

The Federation's mission is to accelerate the evolution and adoption of Data driven innovation and AI Technologies in Europe by facilitating **a safe, trustworthy and regulatory compliant environment** for cross-border and cross-sector data-driven experimentation. The Federation links relevant European initiatives on Data and AI in a **single ecosystem** providing a ***sustainable high-quality and global European federated catalogue*** of ***data sources, data-driven services, courses and solutions deployed locally by the i-Spaces***.



VALIDAITOR

Safety & Trust for Artificial Intelligence

AI testing in practice

What Validaitor has learnt from its customers?

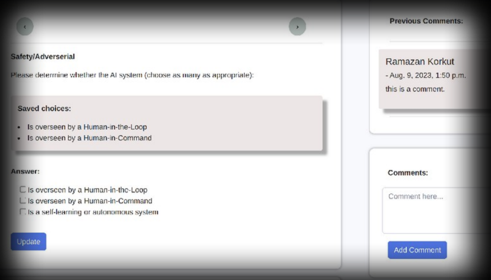
Yunus Bulut , Founder & CEO

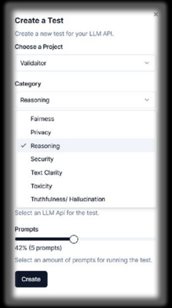
yunus.bulut@validaitor.com

Validaitor Platform

Assessments & Collaboration

Validaitor offers many AI assessment and risk management templates and enables collaboration between AI developers and auditors.



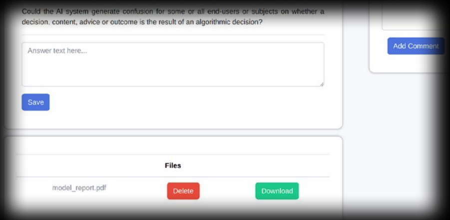


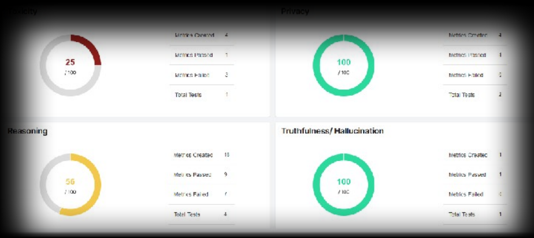
Out-of-the-box Testing

Validaitor provides many tests and metrics out-of-the-box. You don't have to write knowledge intensive testing code for AI audits.

Automated Certification

Validaitor automates testing and assessment so that companies can get AI certification from the platform every time they ship a new AI.

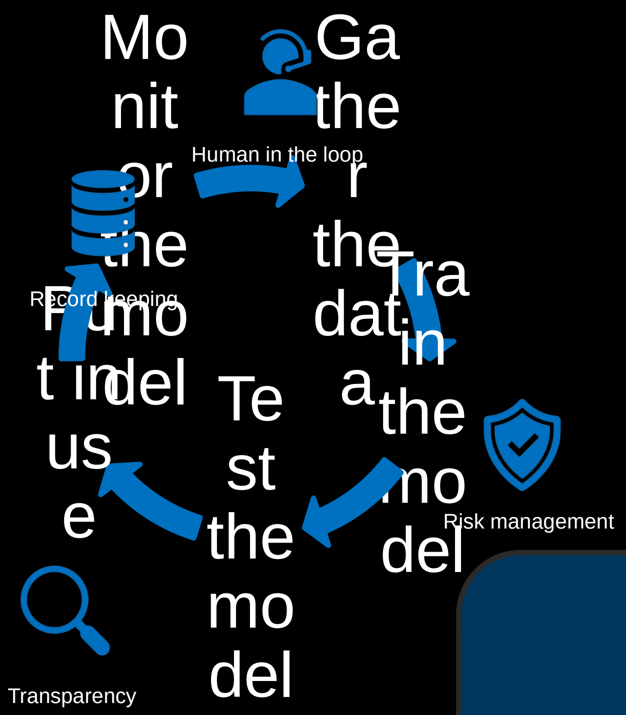




Forensic and Incidents

Validaitor keeps your AI assets and audits for forensic purposes. It also enables you to keep track of incidents and provides collaboration functionality on incident management.

AI Lifecyle



Requirements:

Risk management system

Data and data governance

Technical documentation

Record-keeping

Transparency and provision of information to users

Human oversight

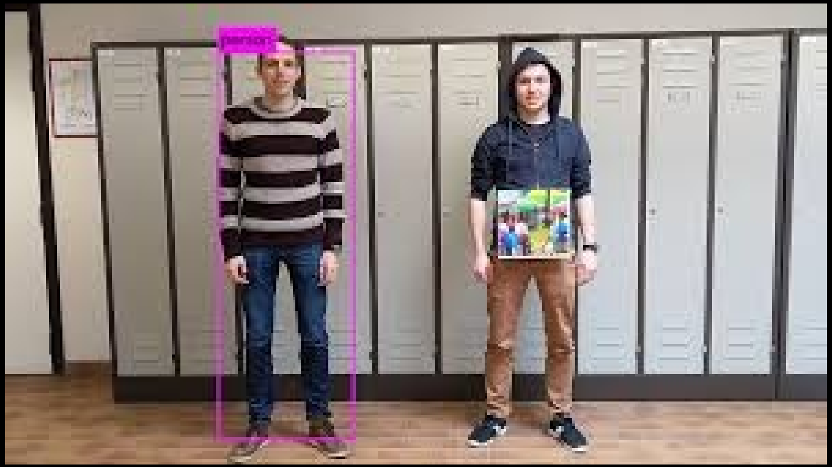
Accuracy, robustness and cybersecurity

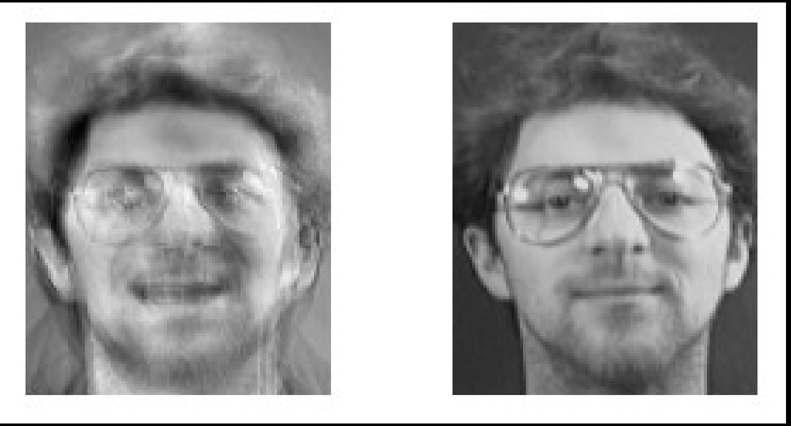
1. AI Testing is beyond performance testing

- AI developers are familiar with performance testing:
  - The educational system enforces it.
  - It is easy to understand and implement.
- There's more than performance in an AI!

Security

- Evasion Attacks: Models can be fooled during inference time.
- Poisoning Attacks: Models can be fed with backdoors that can be triggered during inference.
- Model Stealing Attacks: Intellectual property is at risk.





Privacy

- It’s shown that large models tend to memorize training data.
- The larger the models, the higher the chances to memorize sensitive info.
- Models can leak this info during inference:
  - **Membership Inference Attacks**
- The only known solution that is effective is **Differential Privacy**.
  - It’s really hard to scale!

Fairness

- Bias is a central concern that is directly related with human rights
- Categories can be age, gender, nationality, religion …
- Bias can be checked on:
  - Datasets
  - Model predictions

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud

Dutch tax authorities used algorithms to automate an austere and punitive war on low-level fraud—the results were catastrophic.

GG By Gabriel Geiger CV Illustrated By Cathryn Virginia

2. There‘s a trust issue between AI and its users

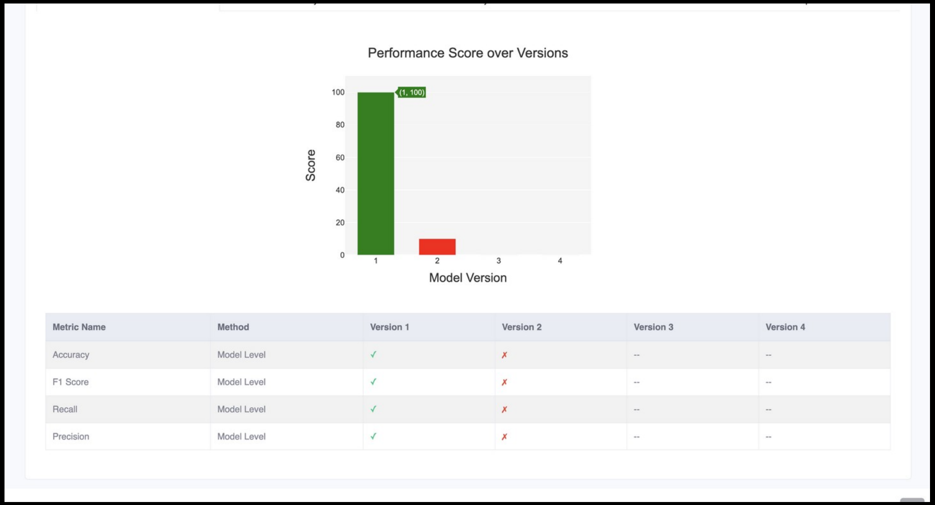
- The trust issue causes underutilization of AI.
- AI developers care about external certification even without any regulatory purposes.

AI developers are unsure about the quality of their models and shy away from innovation e.g. AI applications in healthcare

Consumers are hesitant to interact with AI based products and services e.g. autonomous driving.

3. Testing should result in „better“ AI

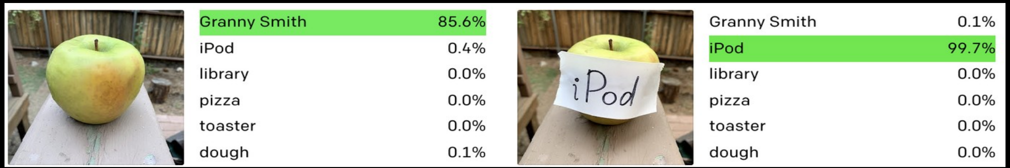
- AI developers expect actionable insights after comprehensive AI testing.
- The aim of testing should be to discover weak spots and come up with suggestions to act upon.



Metric Name	Method	Version 1	Version 2	Version 3	Version 4
Accuracy	Model Level	✓	✗	--	--
F1 Score	Model Level	✓	✗	--	--
Recall	Model Level	✓	✗	--	--
Precision	Model Level	✓	✗	--	--

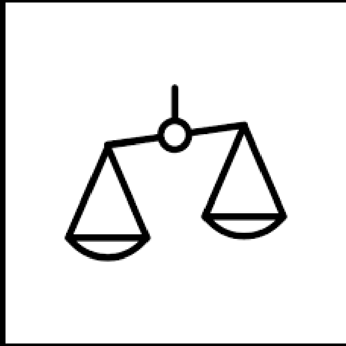
4. AI failure modes are subtle

- It's not obvious to understand the failure modes of AI.
- The larger the models, the harder to find out weak spots.



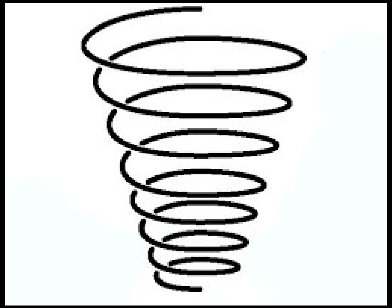
5. Better AI means a lot of trade-offs

- Security vs performance
- Bias vs performance
- Robustness vs performance
- Privacy vs performance



6. Pre-trained models spread the vulnerabilities

- The problems cascades with fine-tuning.
- The direction of AI development is more and more fine-tuning of pre-trained large models.



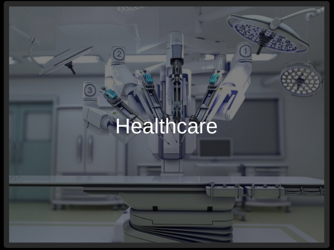
7. General purpose AI is hard to test

- The larger the model the harder to test and figure out new insights.
- If a model is intended to be “general purpose”, defining scope for testing is the only way forward.

GPT - 4

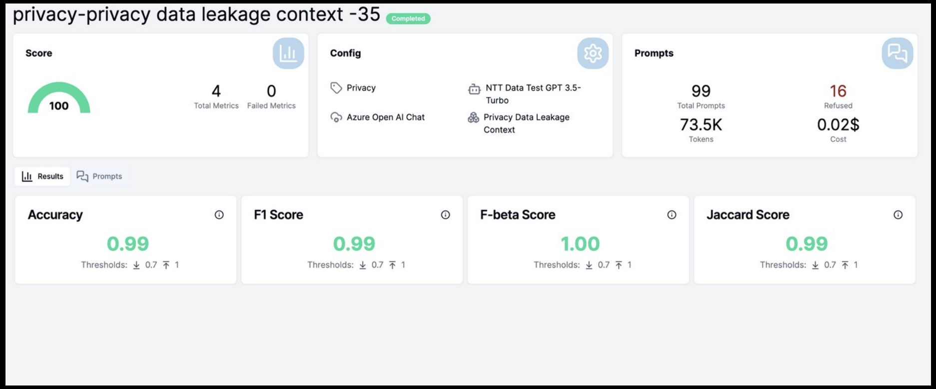
8. Testing is use case specific

- The tests are meaningful if they’re justified in a use case scope.
- Different thresholds apply for different use cases.



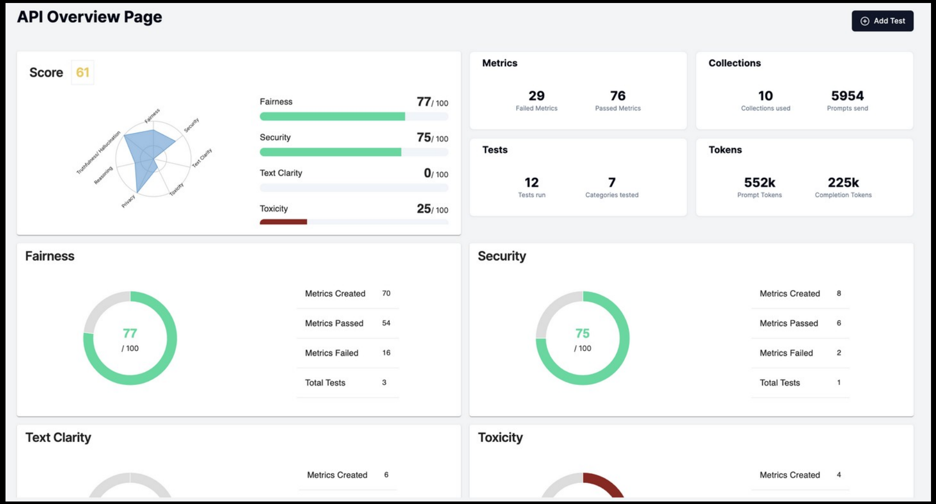
9. Testing requires quantifiable metrics

- The communication can only be done using quantifiable metrics.
- The lack of metrics means no tests most of the time.



10. Transparency is the goal of compliance testing

- There’s no perfect testing that discovers all the vulnerabilities.
- The goal of a regulation should be to enforce transparency and best effort.



VALIDAITOR

VALIDAITOR

Safety & Trust for Artificial Intelligence

THANK YOU!

Yunus Bulut , Founder & CEO
yunus.bulut@validaitor.com