Tools für Good Governance auf der Basis von Normen und Standards als Schlüssel für vertrauenswürdige KI

Dr Jochen Friedrich
Technical Relations Executive

jochen@de.ibm.com

EU AI Act

The AI Act has been in force since August 1, 2024.

It regulates AI technologies on three different levels and applying at different points in time.

Prohibited practices

Applies 6 months after AI Act coming into force, i.e. February 2, 2025.

General purpose AI

Applies 12 months after AI Act coming into force, i.e. August 2, 2025.

AI systems in high-risk areas

For areas listed in Annex III: applies 24 months after AI Act coming into force, i.e. August 2, 2026.

For areas related to Annex I: applies 36 months after AI Act coming into force, i.e. August 2, 2027.

AI Act – EU technical regulation

The AI Act is in its essence a typical EU safety regulation.

IBM has decades-long experience in working with such regulations in the field of hardware compliance.

Basic Principle
(“EU New Legislative Framework”):

Legal acts lay down the essential requirements and define safety objectives.

Harmonised European Standards define the technical way how to fulfil the legal requirements and be compliant with the safety objectives.

Compliance is mandatory for market access.

Harmonised standards are developed in one or more of the European standardisation organisations and are based on formal EU standardisation requests.

AI Act: The Role of Standards

Article 40 establishes the processes of the EU New Legislative Framework to be used.

Article 40

Harmonised standards and standardisation deliverables

1. High-risk AI systems or general-purpose AI models which are in conformity with harmonised standards or parts thereof the references of which have been published in the Official Journal of the European Union in accordance with Regulation (EU) No 1025/2012 shall be presumed to be in conformity with the requirements set out in Section 2 of this Chapter or, as applicable, with the obligations set out in of Chapter V, Sections 2 and 3, of this Regulation, to the extent that those standards cover those requirements or obligations.

Jacques Delors

* 20. Juli 1925
† 27. Dezember 2023

Former president of the European Commission (1985-1995)

Major role in overcoming the “europsclerosis” of the 1980s and accelerating the process of strengthening European unity (EEA, Maastricht treaty, re-organising the European Commission

Introduction of the “New Approach” – later developed in to the New Legislative Framework.

EU Technical Regulation – The New Approach/ New Legislative Framework: Before and After

BEFORE

All technical requirements were part of the legal acts.

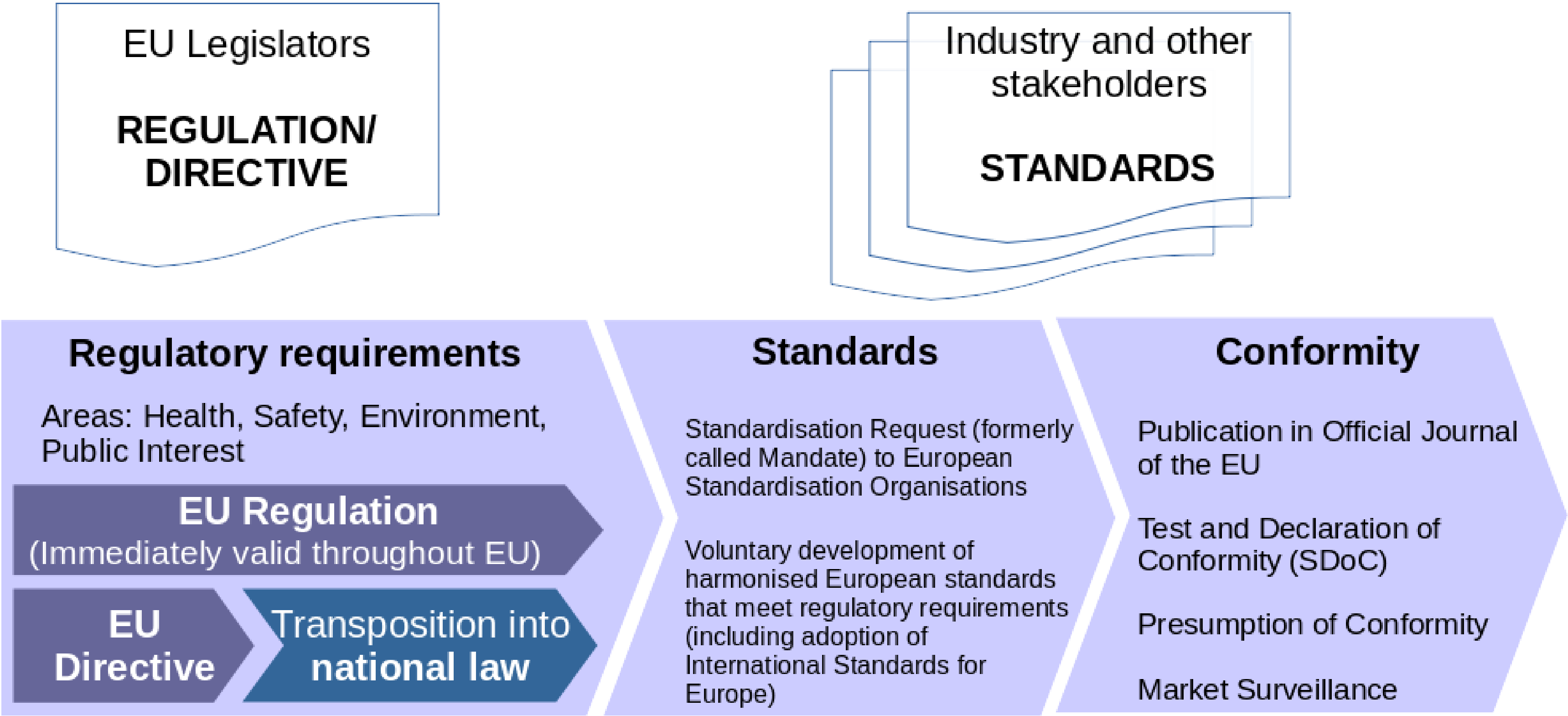
AFTER

Legal acts lay down the legal requirements and define safety objectives.

Harmonised European Standards developed by the private sector define the technical way how to fulfil the legal requirements and be compliant with the safety objectives.

Success Story for Europe:
New Approach – New Legislative Framework (NLF)

- Key element of the EU regulatory framework
- Key element of the EU single market
- Definition of safety objectives in legal acts, technical details laid down in standards
- Compliant products may be brought to the market in the EU under the presumption of conformity
- Innovation friendly – innovation via standards



EU Legislators

REGULATION/ DIRECTIVE

Industry and other stakeholders

STANDARDS

Regulatory requirements

Areas: Health, Safety, Environment, Public Interest

EU Regulation (Immediately valid throughout EU)

EU Directive

Transposition into national law

Standards

Standardisation Request (formerly called Mandate) to European Standardisation Organisations

Voluntary development of harmonised European standards that meet regulatory requirements (including adoption of International Standards for Europe)

Conformity

Publication in Official Journal of the EU

Test and Declaration of Conformity (SDoC)

Presumption of Conformity

Market Surveillance

GPAI: Compliance

Articles 53 and 55 outline how to demonstrate compliance of GPAI models:

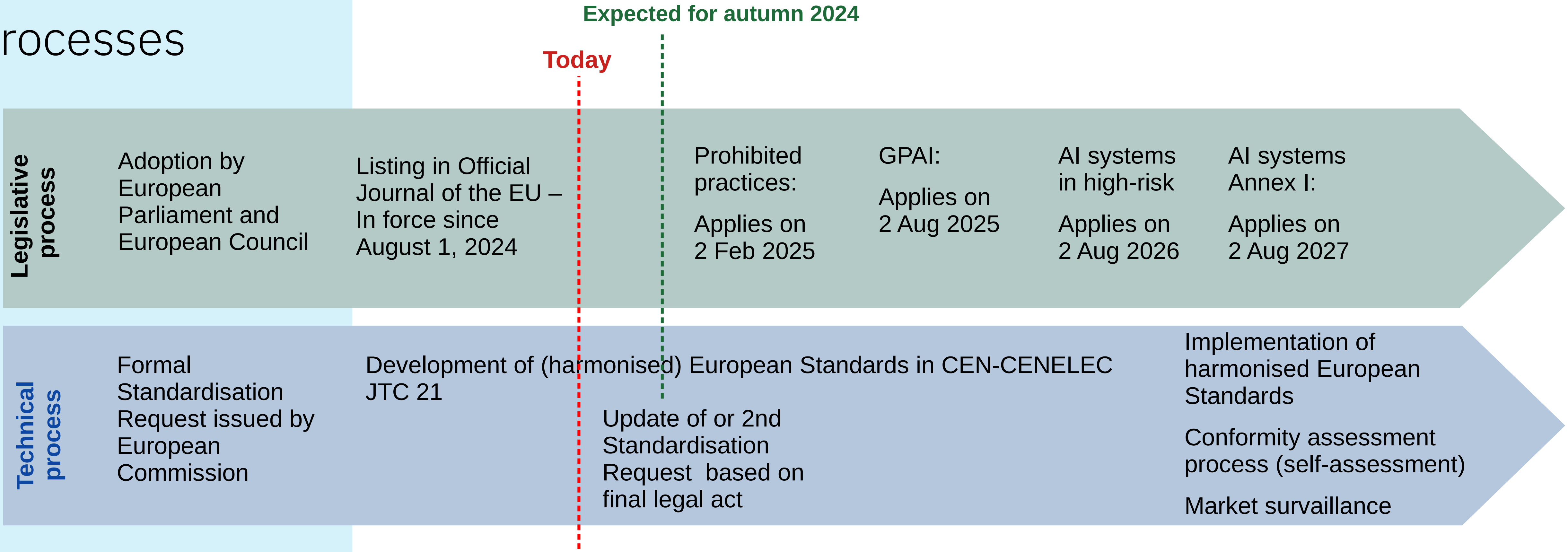
Initially, to bridge the transition time, codes of practice will be provided.

Later on, harmonised standards provide for presumption of conformity.

Providers of general-purpose AI models may rely on codes of practice within the meaning of Article 56 to demonstrate compliance with the obligations set out in paragraph 1 of this Article, until a harmonised standard is published. Compliance with European harmonised standards grants providers the presumption of conformity to the extent that those standards cover those obligations. Providers of general-purpose AI models who do not adhere to an approved code of practice or do not comply with a European harmonised standard shall demonstrate alternative adequate means of compliance for assessment by the Commission.

Providers of general-purpose AI models with systemic risk may rely on codes of practice within the meaning of Article 56 to demonstrate compliance with the obligations set out in paragraph 1 of this Article, until a harmonised standard is published. Compliance with European harmonised standards grants providers the presumption of conformity to the extent that those standards cover those obligations. Providers of general-purpose AI models with systemic risks who do not adhere to an approved code of practice or do not comply with a European harmonised standard shall demonstrate alternative adequate means of compliance for assessment by the Commission.

AI Act: Legislative and technical processes



Expected for autumn 2024

Today

Legislative process

Adoption by European Parliament and European Council

Listing in Official Journal of the EU – In force since August 1, 2024

Prohibited practices:

Applies on 2 Feb 2025

GPAI:

Applies on 2 Aug 2025

AI systems in high-risk

Applies on 2 Aug 2026

AI systems Annex I:

Applies on 2 Aug 2027

Technical process

Formal Standardisation Request issued by European Commission

Development of (harmonised) European Standards in CEN-CENELEC JTC 21

Update of or 2nd Standardisation Request based on final legal act

Implementation of harmonised European Standards

Conformity assessment process (self-assessment)

Market survaillance

GPAI:

Code(s) of practice to bridge time until harmonised standards are available

General Purpose AI (GPAI)

Chapter V, Articles 51ff., address GPAI models

Classification (Article 51):

1. A general-purpose AI model shall be classified as a general-purpose AI model with systemic risk if it meets any of the following conditions:

(a) it has high impact capabilities evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks;

(b) based on a decision of the Commission, ex officio or following a qualified alert from the scientific panel, it has capabilities or an impact equivalent to those set out in point (a) having regard to the criteria set out in Annex XIII.

2.A general-purpose AI model shall be presumed to have high impact capabilities pursuant to paragraph 1, point (a), when the cumulative amount of computation used for its training measured in floating point operations is greater than 10^{25} .

General Purpose AI: Reporting Requirements

Annex XI defines information to be made available

/ general description of GPAI model (e.g. tasks the model is intended to perform, date of release, architecture and number of parameters, …)

/ detailed description of the elements of the model referred to in point 1, and relevant information of the process for the development, e.g. including

// design specifications of the model and training process

// information on the data used for training, testing and validation

Additional information to be provided for GPAI with systemic risk, e.g.:

/ detailed description of the evaluation strategies, including evaluation results, on the basis of available public evaluation protocols and tools or otherwise of other evaluation methodologies. Evaluation strategies shall include evaluation criteria, metrics and the methodology on the identification of limitations.

…

AI systems in high-risk areas

Applies by August 2, 2026

Essential requirements laid down in the AI Act

Requirements have been fairly stable since publication of the first draft of the legal document by the European Commission

Work on harmonised European standards in progress in CEN-CENELEC JTC 21

Work on harmonised European standards

Essential requirements (in the law)

Risk management system

Data and data governance

Technical documentation

Record-keeping

Transparency and provision of information to users

Human oversight

Accuracy, robustness and cybersecurity

EU Standardisation Request in place

Risk management system for AI systems

Governance and quality of datasets used to build AI systems

Record keeping through built-in logging capabilities in AI systems

Transparency and information to the users of AI systems

Human oversight of AI systems

Accuracy specifications for AI systems

Robustness specifications for AI systems

Cybersecurity specifications for AI systems

Quality management system for providers of AI system, including post-market monitoring process.

Conformity assessment for AI systems

Expected in update of Standardisation Request based on final AI Act

Sustainable AI:

”reporting and documentation processes to improve AI systems resource performance, such as reduction of energy and other resources consumption of the high-risk AI system during its lifecycle, and on energy efficient development of general-purpose AI models”

European standards supporting the transparency obligations for GPAI

In the hands of the private sector

With the listing of the AI Act in the Official Journal of the EU – the legal process is done.

Development of the harmonised European Standards required for compliance is in the hands of the private sector.

European Commission: Right of initiative for legal acts

European Council and European Parliament: Two chambers negotiating and agreeing on final legal act

LEGISLATIVE PROCESS COMPLETED WITH LISTING IN OFFICIAL JOURNAL OF EU

PRIVATE SECTOR ENTITLED TO DEVELOP HARMONISED EUROPEAN STANDARDS

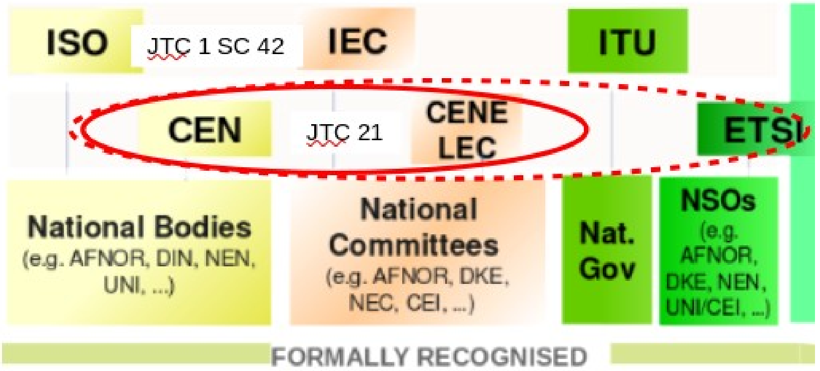
Industry

Civil Society

Research

Academia

...



CONFORMITY ASSESSMENT

Option A: Build up test lab for self-assessment internally

Option B: Work with notified bodies

EU MEMBER STATES

Accreditation of notified bodies

Building up of market surveillance

In the hands of the private sector

With the listing of the AI Act in the Official Journal of the EU – the legal process is done.

Development of the harmonised European Standards required for compliance is in the hands of the private sector.

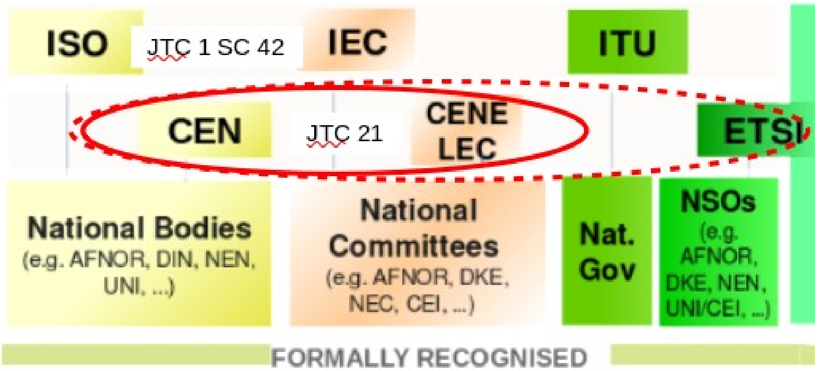
European Commission: Right of initiative for legal acts

European Council and European Parliament: Two chambers negotiating and agreeing on final legal act

LEGISLATIVE PROCESS COMPLETED WITH LISTING IN OFFICIAL JOURNAL OF EU

PRIVATE SECTOR ENTITLED TO DEVELOP HARMONISED EUROPEAN STANDARDS

- Industry
- Civil Society
- Research
- Academia
- …



CONFORMITY ASSESSMENT

Option A: Build up test lab for self-assessment internally

Option B: Work with notified bodies

EU MEMBER STATES

Accreditation of notified bodies

Building up of market surveillance

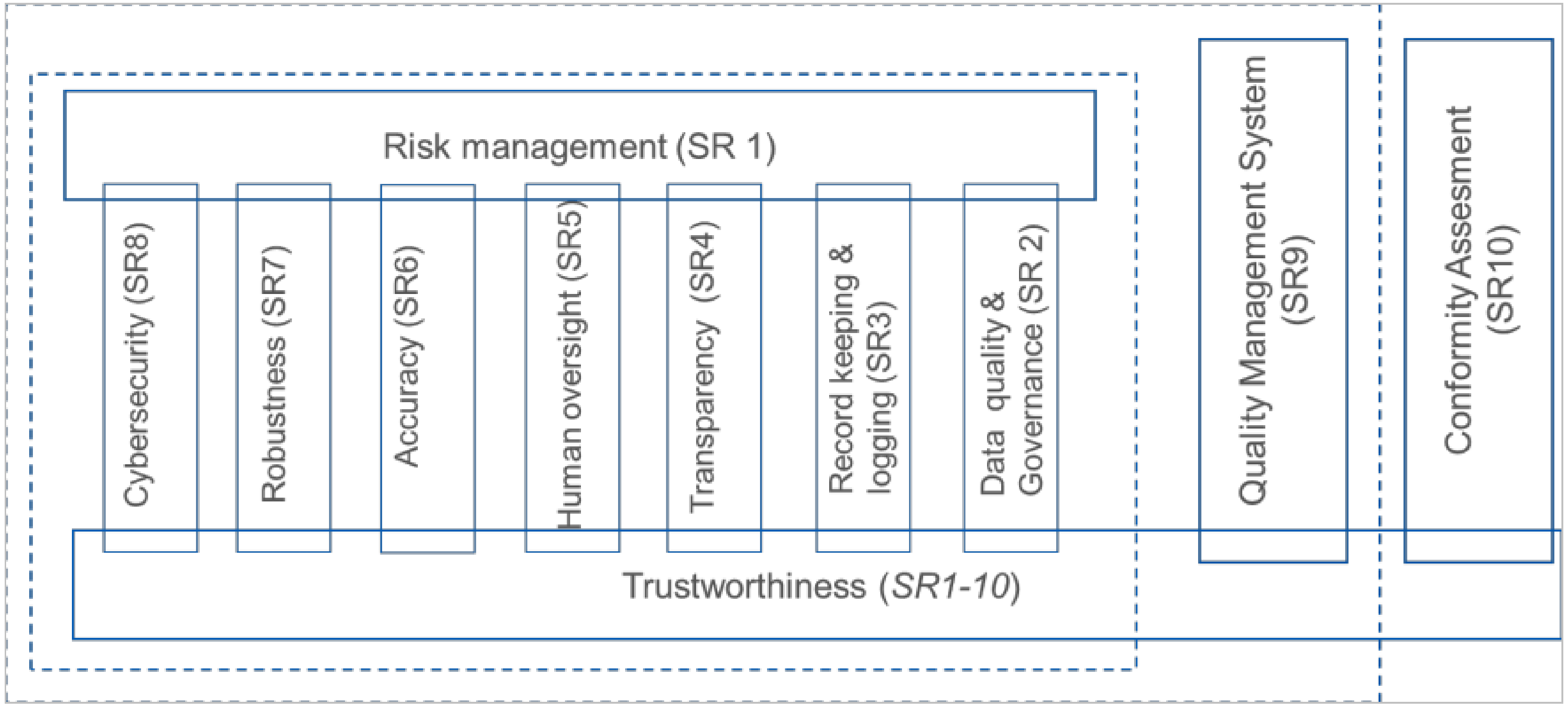
Harmonised European Standards for AI

Adopt ISO standards as European Standards whenever possible.

Close gaps regarding the requirements with the AI Act according to the architecture of standards below.

Provide the overall AI Trustworthiness Framework outlining the standards (and profiles) needed for compliance with the AI Act.

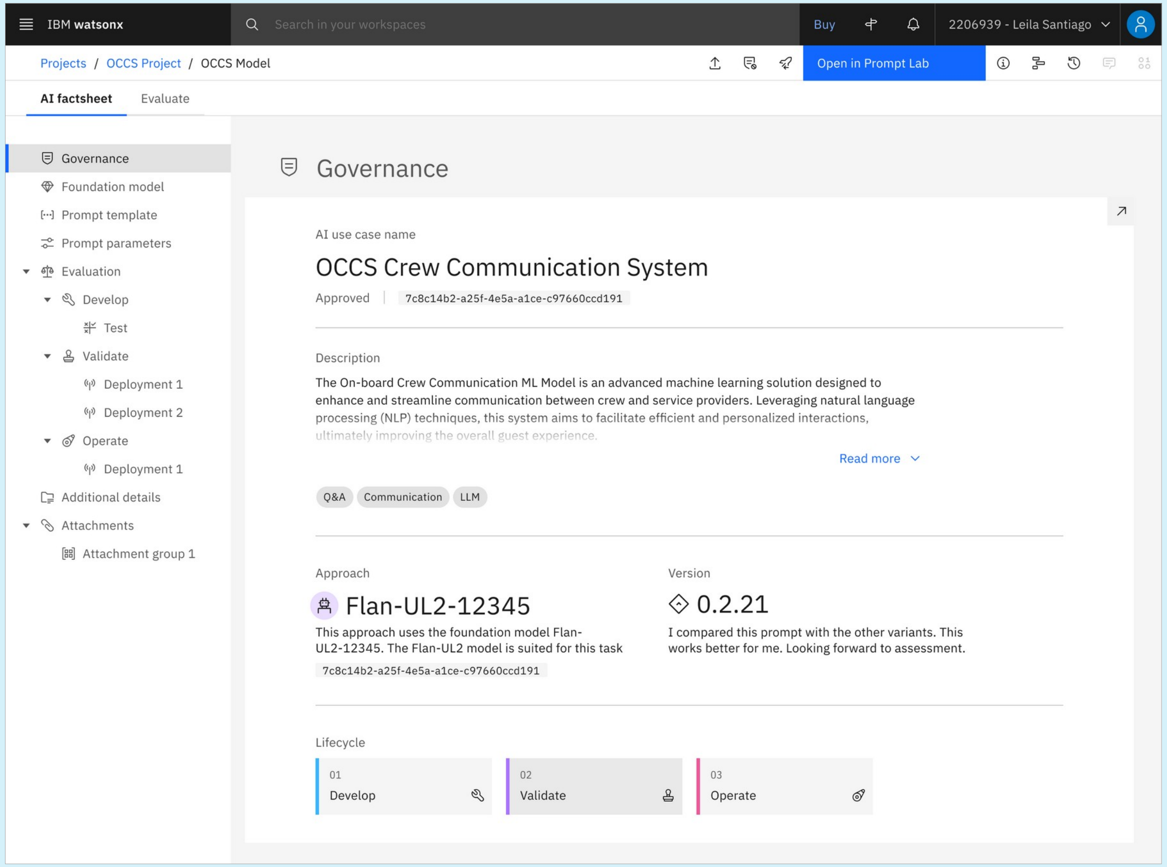
Standards intended to support the AI Act will be mapped on the following architecture of standards.



What IBM offers

Documentation: Your logbook for AI

- Document use cases across the AI lifecycle
- Automate the capture of model metadata
- Review test results, changes, licenses and more
- Integrated reporting and factsheet export



Governance

AI use case name

OCCS Crew Communication System

Approved | 7c8c14b2-a25f-4e5a-a1ce-c97660ccd191

Description

The On-board Crew Communication ML Model is an advanced machine learning solution designed to enhance and streamline communication between crew and service providers. Leveraging natural language processing (NLP) techniques, this system aims to facilitate efficient and personalized interactions, ultimately improving the overall guest experience.

Read more

Q&A Communication LLM

Approach

Flan-UL2-12345

This approach uses the foundation model Flan-UL2-12345. The Flan-UL2 model is suited for this task

7c8c14b2-a25f-4e5a-a1ce-c97660ccd191

Version

0.2.21

I compared this prompt with the other variants. This works better for me. Looking forward to assessment.

Lifecycle

01 Develop

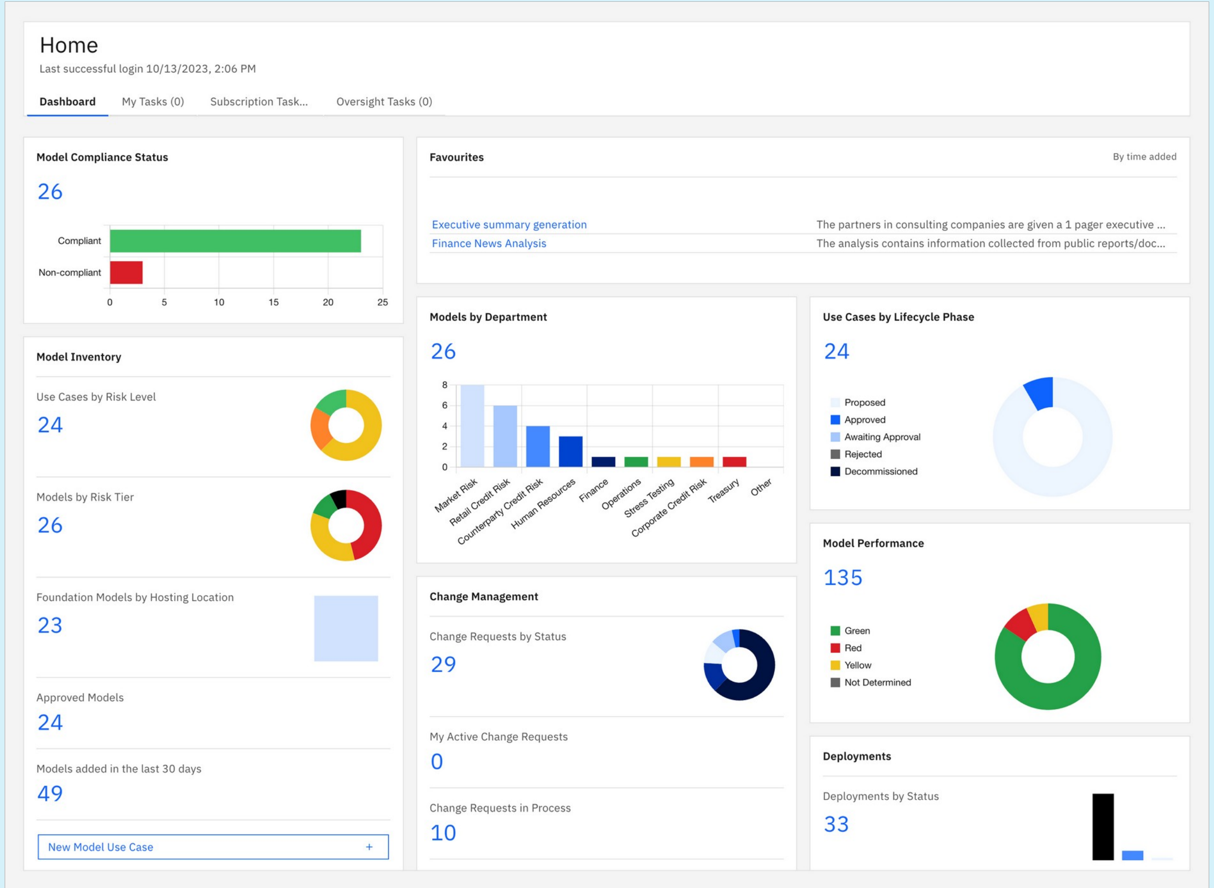
02 Validate

03 Operate

What IBM offers

AI Risk Mgmt: Your AI cockpit.

- EU AI Act Risk Assessment
- Workflows to enforce policies
- Dashboard for any AI on any infrastructure
- End-to-end issue management and task definition
- Monitoring and alerts for threshold violations



Home

Last successful login 10/13/2023, 2:06 PM

Dashboard | My Tasks (0) | Subscription Task... | Oversight Tasks (0)

Model Compliance Status

26

Favourites

By time added

Executive summary generation	The partners in consulting companies are given a 1 pager executive ...
Finance News Analysis	The analysis contains information collected from public reports/doc...

Model Inventory

Use Cases by Risk Level

24

Models by Risk Tier

26

Foundation Models by Hosting Location

23

Approved Models

24

Models added in the last 30 days

49

New Model Use Case +

Models by Department

26

Use Cases by Lifecycle Phase

24

Change Management

Change Requests by Status

29

My Active Change Requests

0

Change Requests in Process

10

Model Performance

135

Deployments

Deployments by Status

33

How does **watsonx.governance** Support clients with the EU AI Act

●

Applicability and Risk Categorisations [Articles 5,6 & 7]

EU AI applicability and Risk Categorisation Assessment Questionnaire

◑

Quality Management System [Article 17]

watsonx.governance as a comprehensive solution acts as a quality management system with workflow, policy, testing, documentation , incident response.

Future releases to ship Policy and control content + specific incident response workflow

●

AI Risk Identification & Assessment [Article 9]

Questionnaire for Risk Identification with use of IBM AI Risk Atlas

Assessment cycles and documentation of controls for all identified risks

●

Technical Documentation & Record Keeping [Articles 11,12 & 18]

Automated Factsheet generation and synchronisation throughout deployment cycles

◑

Accuracy, Robustness & Cyber Monitoring [Article 15]

AI Use Case Monitoring [OpenScale] for accuracy + other metrics

Additonal capability required for Cyber

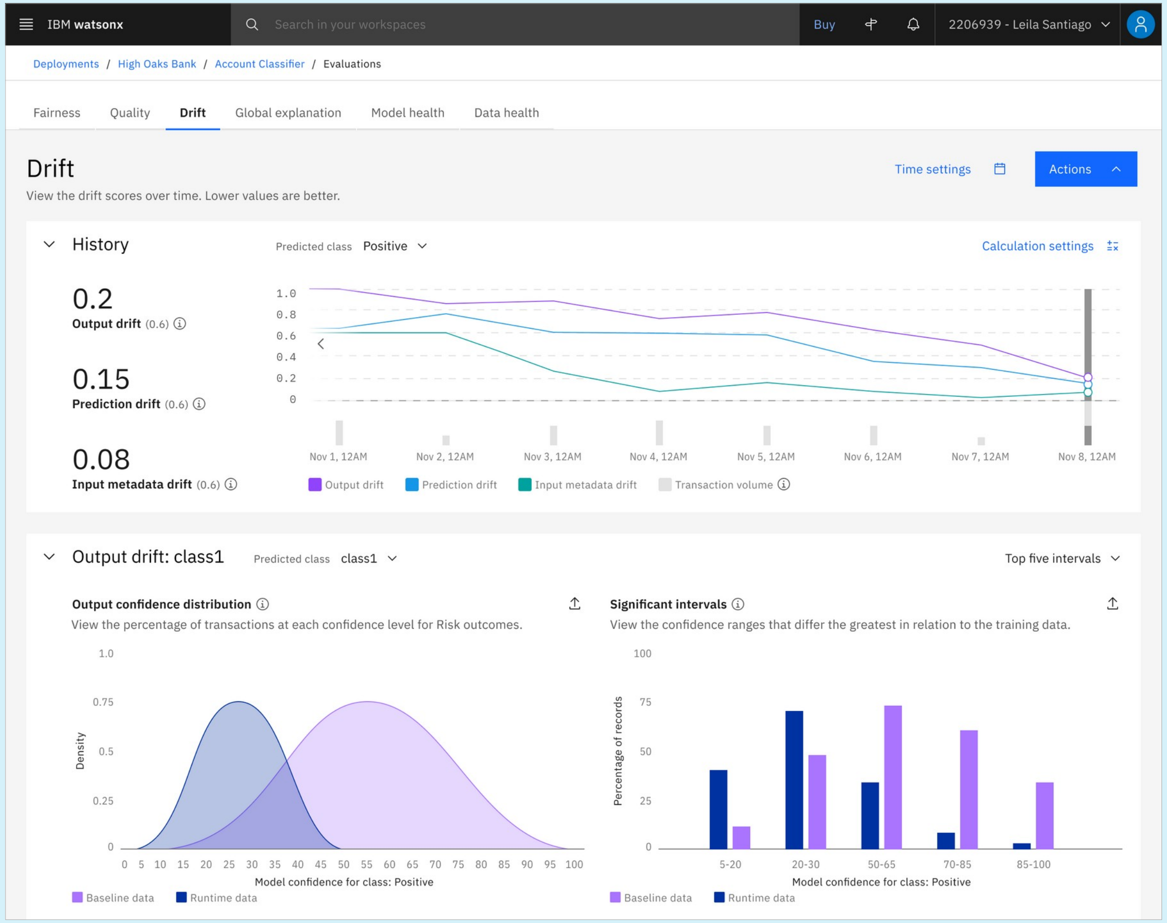
● – *Full support*

◑ – *Partial support*

What IBM offers

AI Evaluation: Monitor performance

- Monitor model accuracy, drift, bias, explainability, quality
- Set thresholds and get alerts when key metrics are breached
- Provide explainable model results in support of audits



Deployments / High Oaks Bank / Account Classifier / Evaluations

Drift

View the drift scores over time. Lower values are better.

History

0.2
Output drift (0.6)

0.15
Prediction drift (0.6)

0.08
Input metadata drift (0.6)

Output drift: class1

Output confidence distribution

View the percentage of transactions at each confidence level for Risk outcomes.

Significant intervals

View the confidence ranges that differ the greatest in relation to the training data.

Trustworthy AI: Key focus for IBM

The EU AI Act is a typical EU safety regulation. Harmonised Europeans standards are key for compliance with the legal requirements.

IBM has decade long experience in operating under the EU New Legislative Framework and providing standards to meet EU safety and security objectives.

IBM has long advocated for trustworthy AI and offers governance tools (watsonx.governance) for managing trustworthy AI and operating with good governance.

IBM actively contributes to standardisation work at international and European level bringing in its expertise and supporting the implementation of the AI Act.

IBM ensures that its AI governance tools and technologies are fully in synch with the harmonised European standards and thus supporting compliance with the EU AI Act.

Many thanks for your attention

Dr. Jochen Friedrich

jochen@de.ibm.com
https://www.linkedin.com/in/jochenfriedrich/