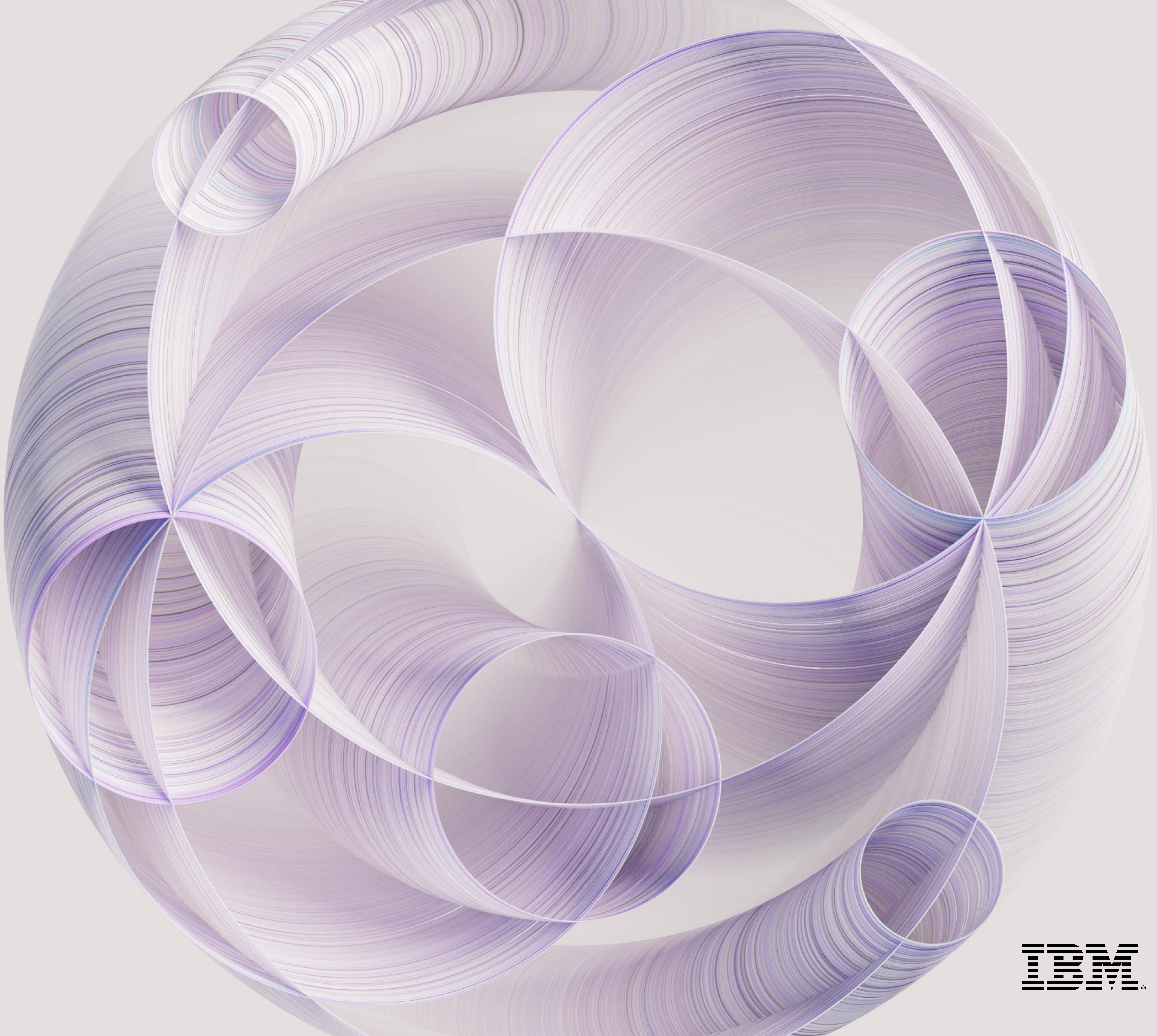
Accelerate responsible, transparent and explainable AI



watsonx.governance

Foundation models are bringing an inflection point in AI…

…but how enterprises adopt and execute will define whether they unlock value at scale

The speed, scope, and scale of generative AI impact is unprecedented

Massive early adoption

80%

of enterprises are working with or planning to leverage foundation models and adopt generative AI

Broad-reaching and deep impact

Generative AI could raise global GDP by

7%

within 10 years

Critical focus of AI activity and investment

Generative AI expected to represent

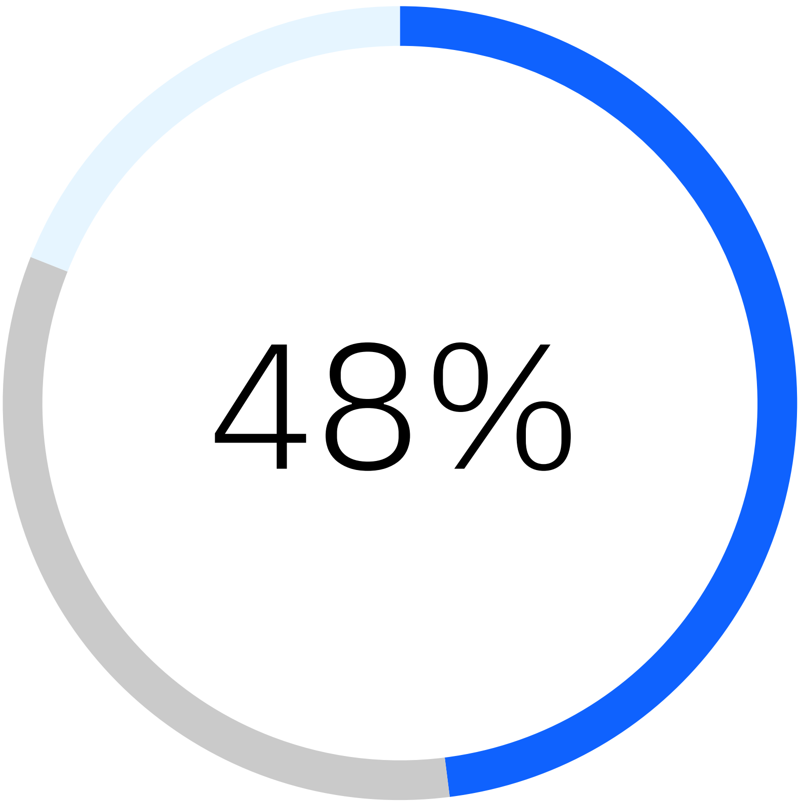
30%

of overall market by 2025

Business leaders face challenges in scaling AI across the enterprise with trust

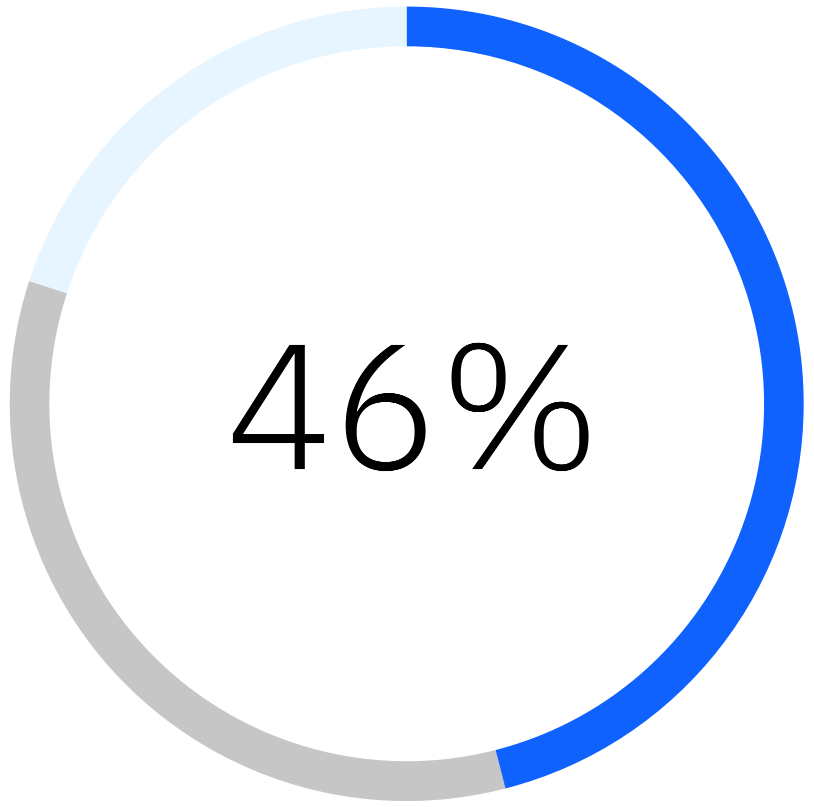
80% of surveyed business leaders see at least one of these ethical issues as a major concern[1]

Explainability



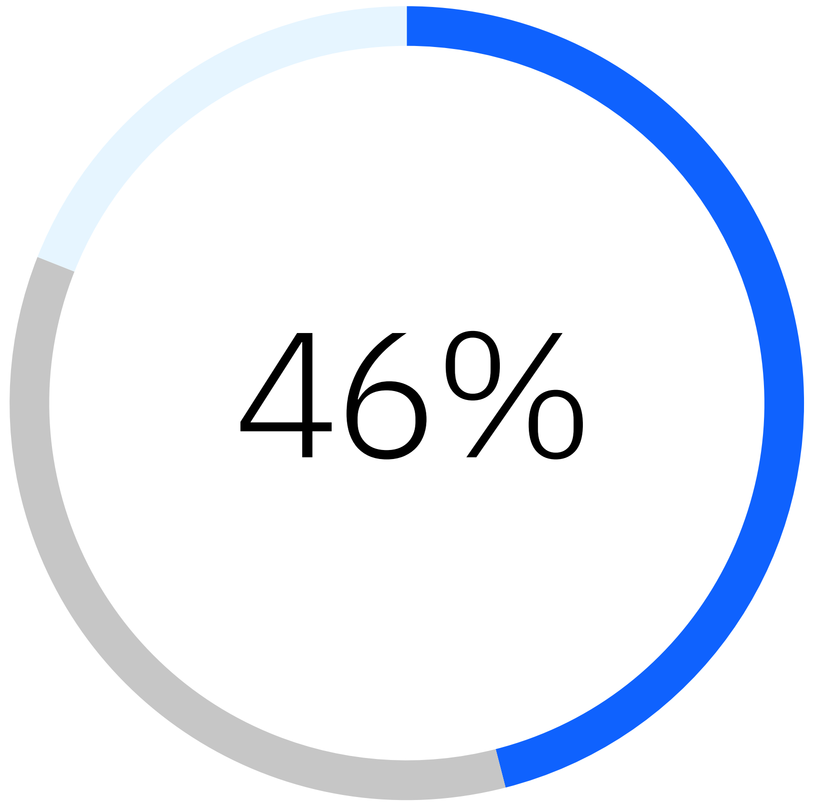
believe decisions made by generative AI are not sufficiently **explainable**

Ethics



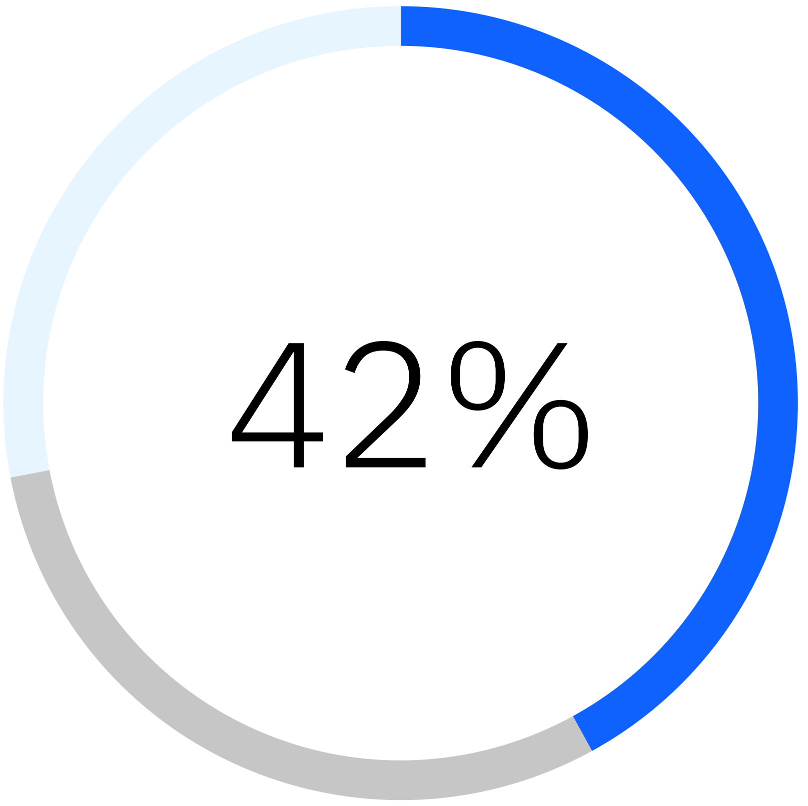
concerned about the safety and **ethical** aspects of generative AI

Bias



believe that generative AI will propagate established **biases**

Trust



believe generative AI cannot be **trusted**

🟦 Agree ▨ Neutral □ Disagree

Foundation model risks

Risk Associated with Input

Training and Fine-tuning Phase

- Bias
- Data poisoning attacks
- Legal restrictions on data
  - Copyright and other IP issues
  - Inclusion of PI and SPI
- Data transparency challenges

Inference Phase

- Disclosure of PI/SPI/Copyright/other IP information as a part of prompt
- Adversarial attacks like evasion, prompt injection, prompt leaking, and jail breaking

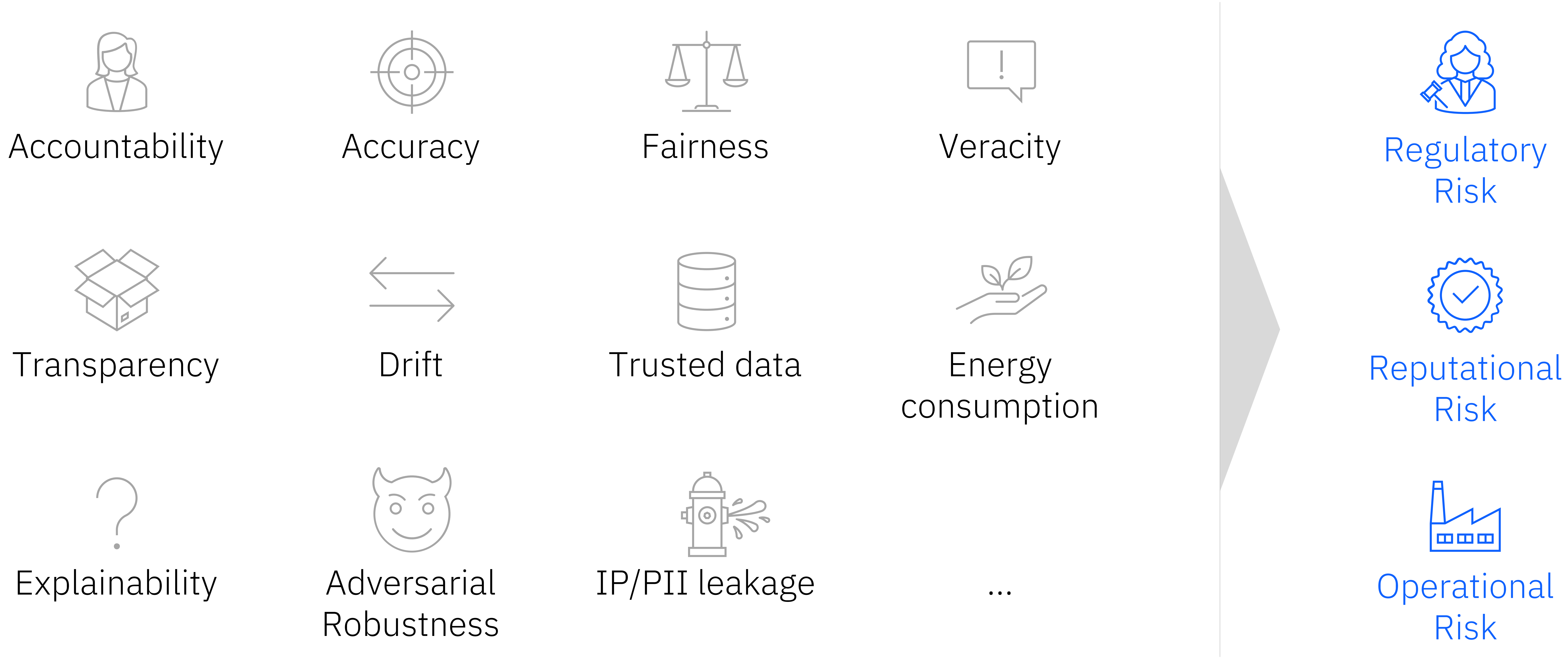
Risk Associated with Output

- Bias in generated content
- Performance disparity
- Copyright infringement
- Value alignment issues (e.g., Hallucination)
- Misuse
- Exposing PI and SPI in the output
- Explainability challenges
- Traceability challenges

Challenges

- Transparency challenges
- Challenge around assigning responsibility
- IP issues
- Human exploitation
- Impact on jobs
- Environmental Impact
- Diversity and Inclusion
- Human agency
- Impact on education

Elements of AI Risk



Accountability

Accuracy

Fairness

Veracity

Transparency

Drift

Trusted data

Energy consumption

Explainability

Adversarial Robustness

IP/PII leakage

...

Regulatory Risk

Reputational Risk

Operational Risk

Risk is everyone's business.

In today's turbulent environment, the need to take on risk with confidence is greater than ever before.

90%

of compliance leaders expect evolving business, regulatory, and customer demands to increase compliance-related operating costs by up to 30%.[1]

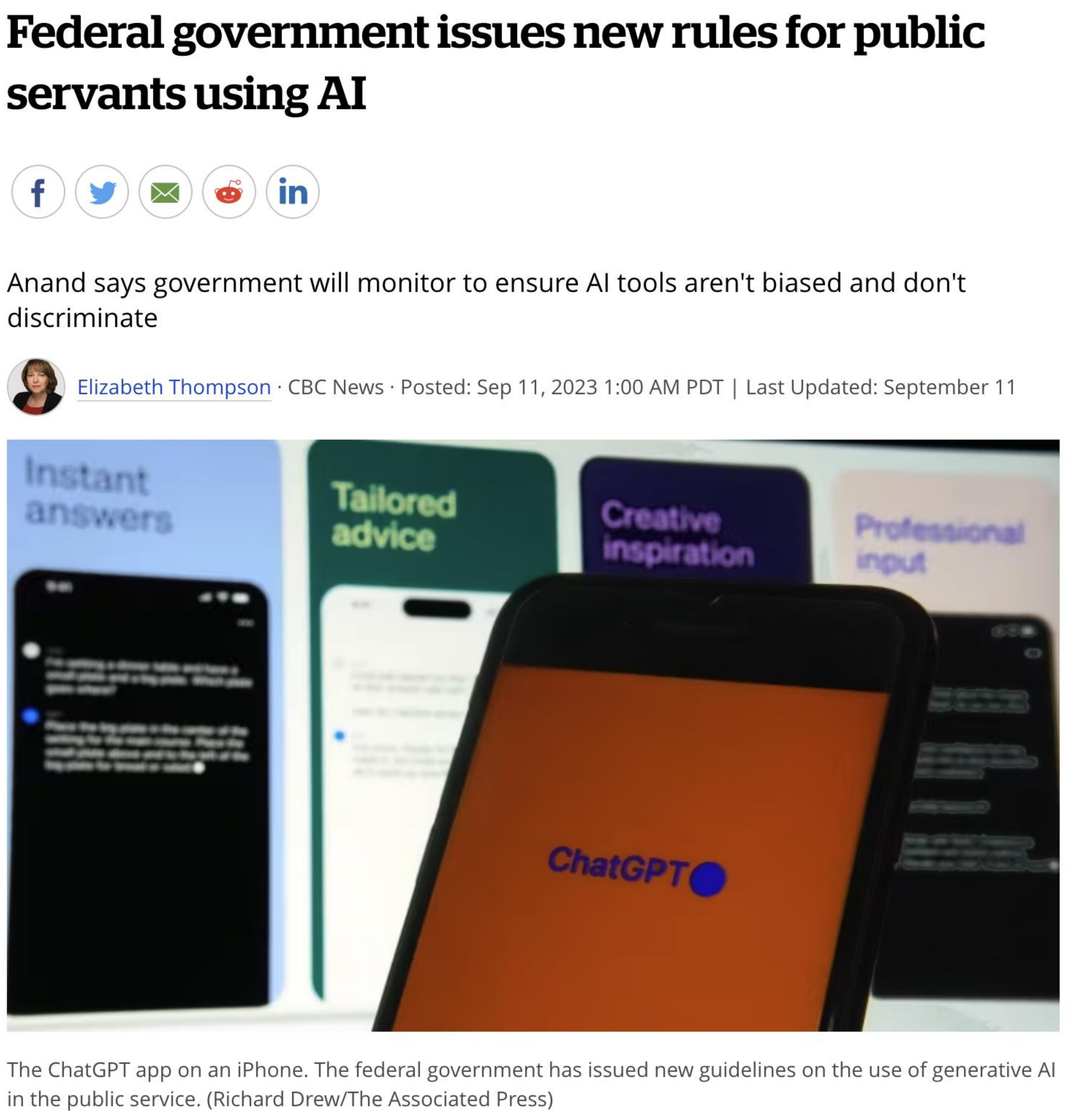
79%

of organizations report that keeping up with the speed of digital and other transformations is a significant risk management challenge.[2]

77%

of organizations recognize the need to upgrade their Third-Party Risk Management operating model.[3]

AI regulations are coming closer

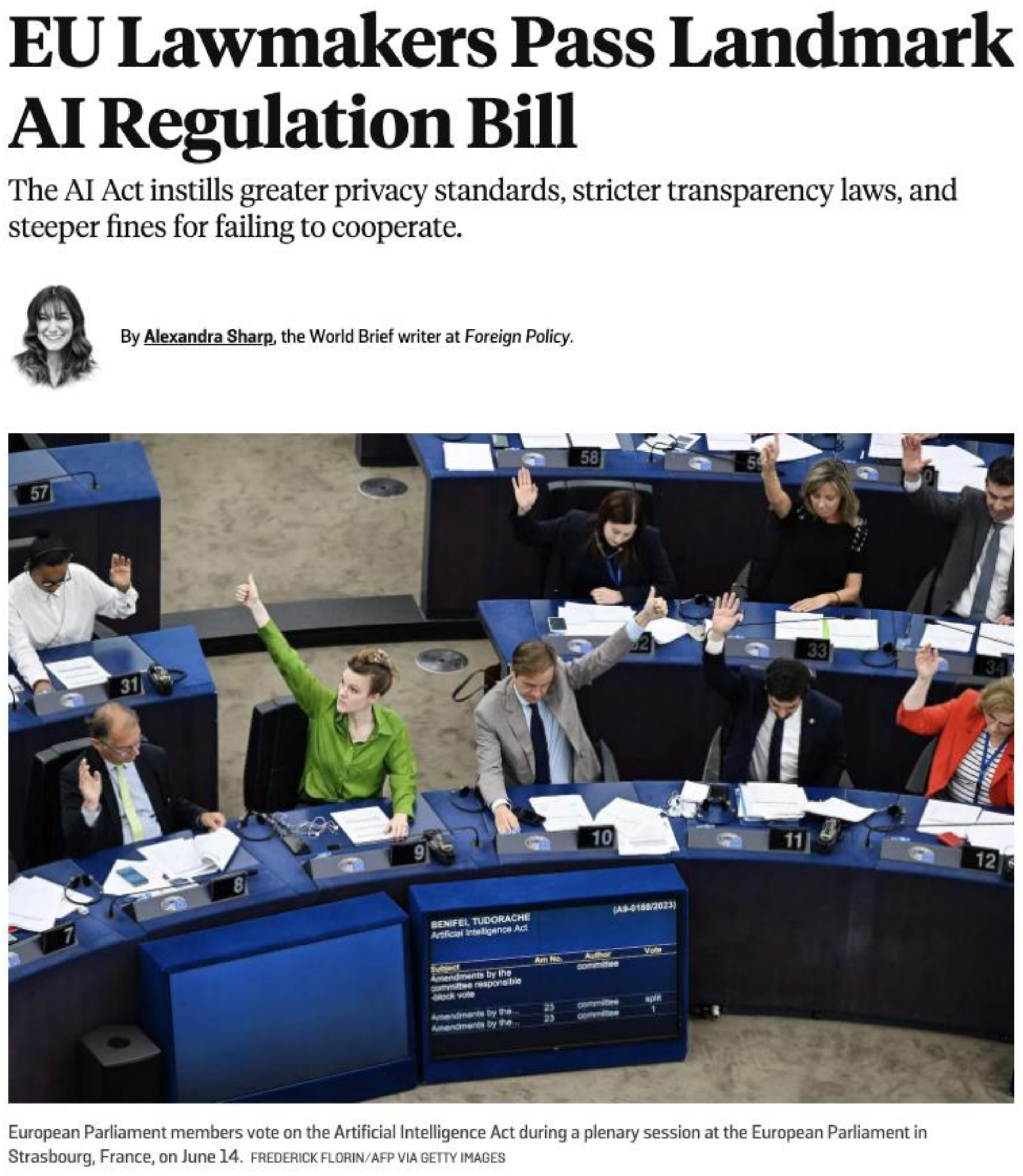


Federal government issues new rules for public servants using AI

Anand says government will monitor to ensure AI tools aren't biased and don't discriminate

Elizabeth Thompson · CBC News · Posted: Sep 11, 2023 1:00 AM PDT | Last Updated: September 11

The ChatGPT app on an iPhone. The federal government has issued new guidelines on the use of generative AI in the public service. (Richard Drew/The Associated Press)

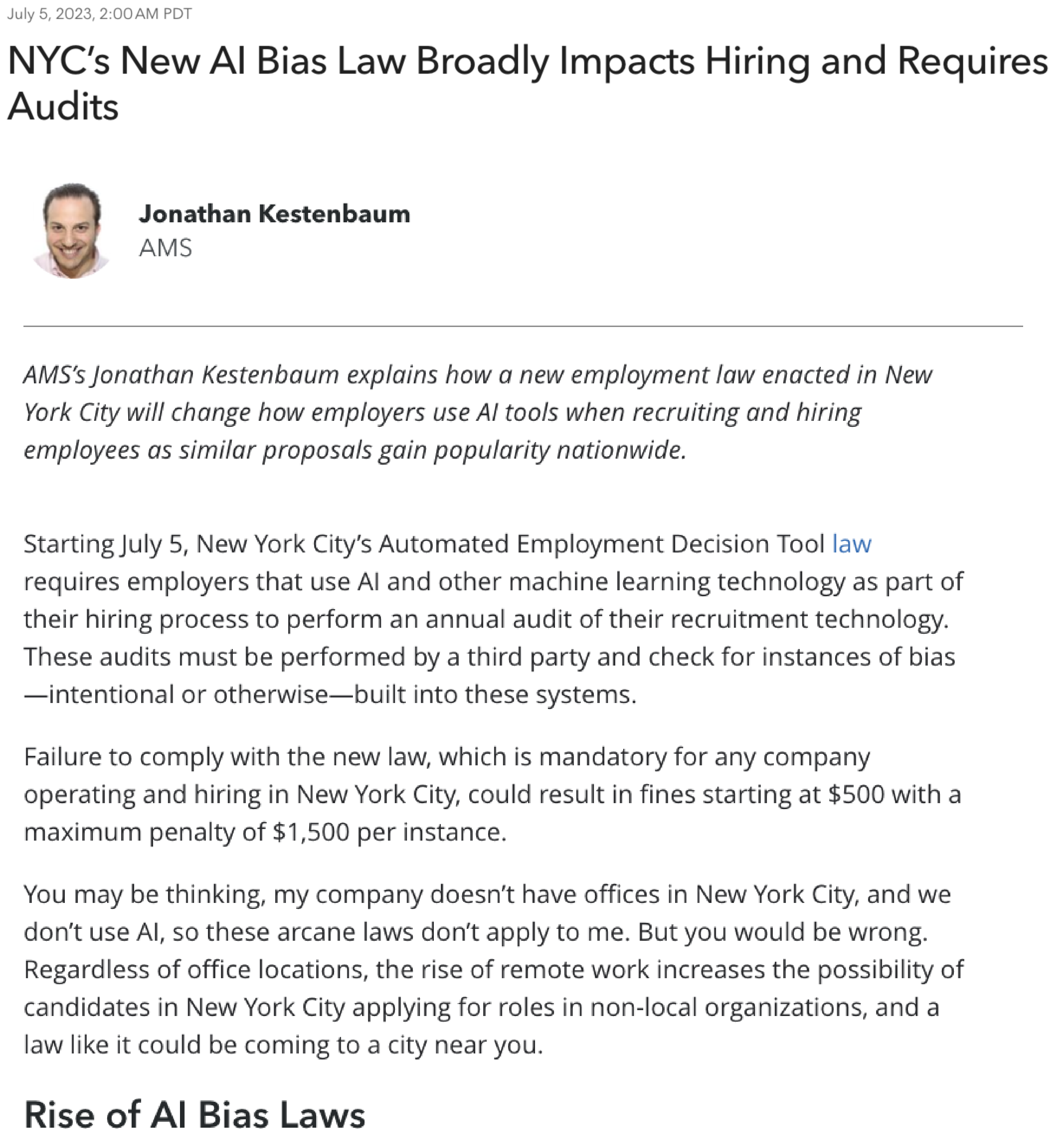


EU Lawmakers Pass Landmark AI Regulation Bill

The AI Act instills greater privacy standards, stricter transparency laws, and steeper fines for failing to cooperate.

By Alexandra Sharp, the World Brief writer at *Foreign Policy*.

European Parliament members vote on the Artificial Intelligence Act during a plenary session at the European Parliament in Strasbourg, France, on June 14. FREDERICK FLORIN/AFP VIA GETTY IMAGES



July 5, 2023, 2:00 AM PDT

NYC's New AI Bias Law Broadly Impacts Hiring and Requires Audits

Jonathan Kestenbaum
AMS

AMS's Jonathan Kestenbaum explains how a new employment law enacted in New York City will change how employers use AI tools when recruiting and hiring employees as similar proposals gain popularity nationwide.

Starting July 5, New York City's Automated Employment Decision Tool law requires employers that use AI and other machine learning technology as part of their hiring process to perform an annual audit of their recruitment technology. These audits must be performed by a third party and check for instances of bias—intentional or otherwise—built into these systems.

Failure to comply with the new law, which is mandatory for any company operating and hiring in New York City, could result in fines starting at $500 with a maximum penalty of $1,500 per instance.

You may be thinking, my company doesn't have offices in New York City, and we don't use AI, so these arcane laws don't apply to me. But you would be wrong. Regardless of office locations, the rise of remote work increases the possibility of candidates in New York City applying for roles in non-local organizations, and a law like it could be coming to a city near you.

Rise of AI Bias Laws

What makes a generative AI platform trustworthy?

How was it trained?

- Garbage in, garbage out
- An enterprise should not use a foundation model trained with a Wikipedia crawl
- The training material must be huge and comprehensive, but must also be curated

Can it detect & minimize bias & hallucination?

- How does the platform detect and correct bias?
- How can it prevent hallucination (providing random and untrue answers with absolute aplomb and conviction)?

Is it transparent?

- Open vs. black-box
- How to audit and explain a model and the answers it generates?
- Does the model track drift and bias? And how does it address them?

Does it support regulatory compliance?

- How do foundation models and their usage comply with privacy and government regulations?
- What are the guardrails?
- Who is responsible for inadvertently exposed personal identifiable information or a “wrong answer”?

Is it safe?

- Who has control over the model, input data, and output data?
- Can you ensure that confidential information is not given out?
- How is it monitored?
- What safety features and guardrails are in place?

Can it be customized?

- Hybrid and multicloud?
- Can the model be fine-tuned with your data?
- Can it be enhanced and extended to make it more suitable for specific use cases?
- How will it integrate with other applications?

AI needs governance – the process of directing, monitoring and managing the AI activities of an organization

Governance necessities

Monitor and evaluate

- Monitor predictive models for fairness, accuracy, and drift
- Monitor generative models for PII and HAP, with additional monitors coming soon
- Explain model predictions and output

Track facts and metrics

- Automatically gather model metrics and metadata
- Provide model information in a fully-managed, searchable catalog
- Track models throughout the entire lifecycle

Manage lifecycle and risk

- Fully customize model approval workflows, from initial request to production deployment
- Track risk for all models across the enterprise
- Configure dashboards and reporting for model performance

* Coming 1Q2024

AI governance is complicated

AI governance collaboration requires lots of **manual work**; amplified by changes in data and model versions.

Companies have models in **multiple tools, applications and platforms**, developed inside and outside the organization

Governance is **not a one-size-fits-all** approach.

The **lack of tools for collaboration and communication** impacts stakeholder management

IBM watsonx.governance

↓

a powerful toolkit built to direct, manage and monitor the AI activities of an organization

Build enduring consumer trust with your brand

Boost productivity and accelerate business outcomes

Mitigate risk and minimize cost of compliance

watsonx.governance

Accelerate responsible, transparent and explainable AI

One unified, integrated AI Governance platform to govern generative AI and predictive ML

Lifecycle Governance

Govern across the AI lifecycle. Automate and consolidate tools, applications and platforms. Capture metadata at each stage and support models built and deployed in 3rd party tools.

Risk Management

Manage risk & protect reputation by automating workflows to ensure quality and better detect bias and drift.

Regulatory Compliance

Adhere to regulatory compliance by translating growing regulations into enforceable policies.

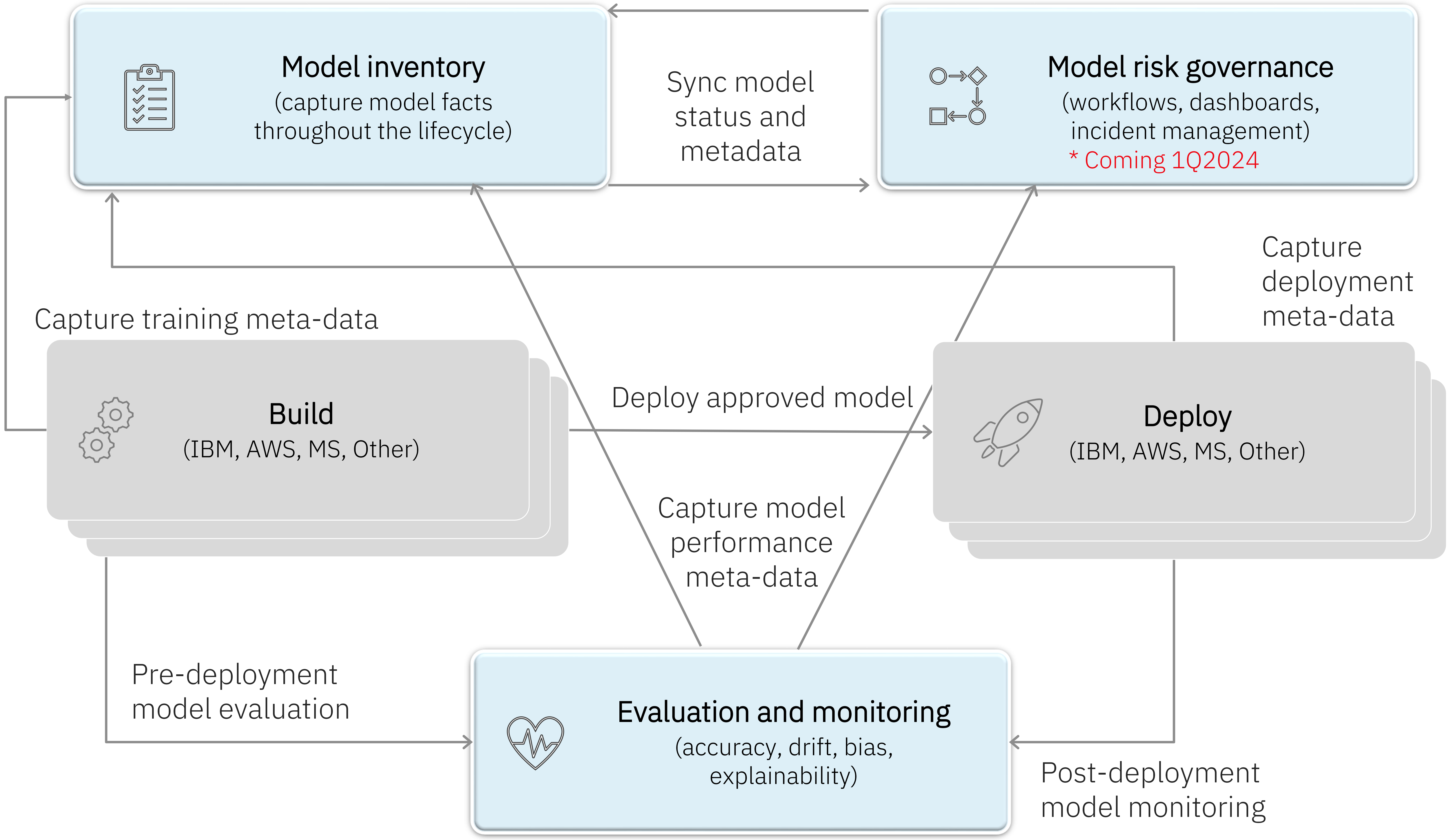
Comprehensive
Govern the end-to-end AI lifecycle with metadata capture at each stage

Open
Support governance of models built and deployed in 3rd party tools.

Automatic metadata recording
and data transformation/lineage capture though Python notebooks.

watsonx.governance

Trusted: Accelerate responsible, transparent, and explainable AI workflows


Model inventory
(capture model facts throughout the lifecycle)
Sync model status and metadata
Model risk governance
(workflows, dashboards, incident management)
* Coming 1Q2024
Capture deployment meta-data
Capture training meta-data
Build
(IBM, AWS, MS, Other)
Deploy approved model
Deploy
(IBM, AWS, MS, Other)
Capture model performance meta-data
Pre-deployment model evaluation
Evaluation and monitoring
(accuracy, drift, bias, explainability)
Post-deployment model monitoring

A toolkit for AI governance

- Govern generative AI and traditional ML models across the entire AI lifecycle
- Automate and consolidate multiple tools, applications & platforms while documenting the origin of data sets, models meta data, and pipelines
- Manage risk and protect reputation by automating workflows to better detect fairness, bias, and drift
- Improve adherence to AI regulations, such as the proposed EU AI Act, and internal compliance standards

watsonx.governance

Manage risk across the enterprise

Risks and governance requirements differ by:

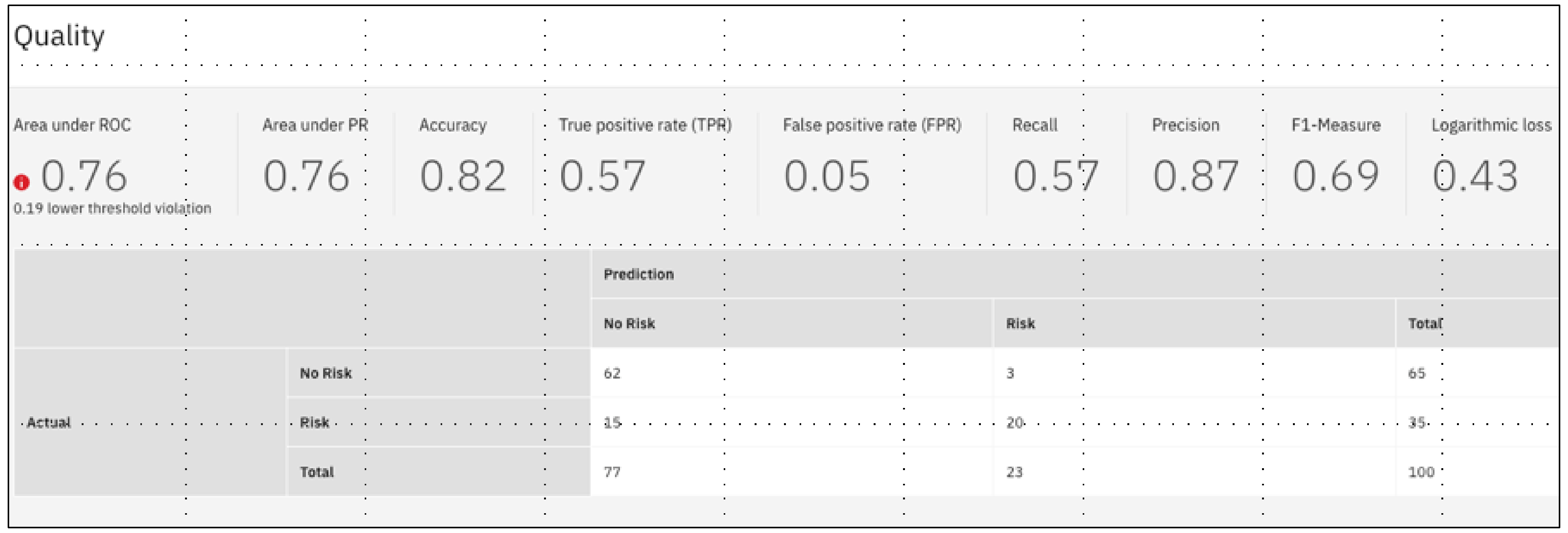
- Use Case
- Industry
- Geography
- Company
- Technology used

Your governance solution needs adjusts to your specific situation:

- Risk assessment
- Governance workflows
- Dashboards
- Model metadata
- Monitoring metrics

watsonx.governance

Manage risk across the enterprise



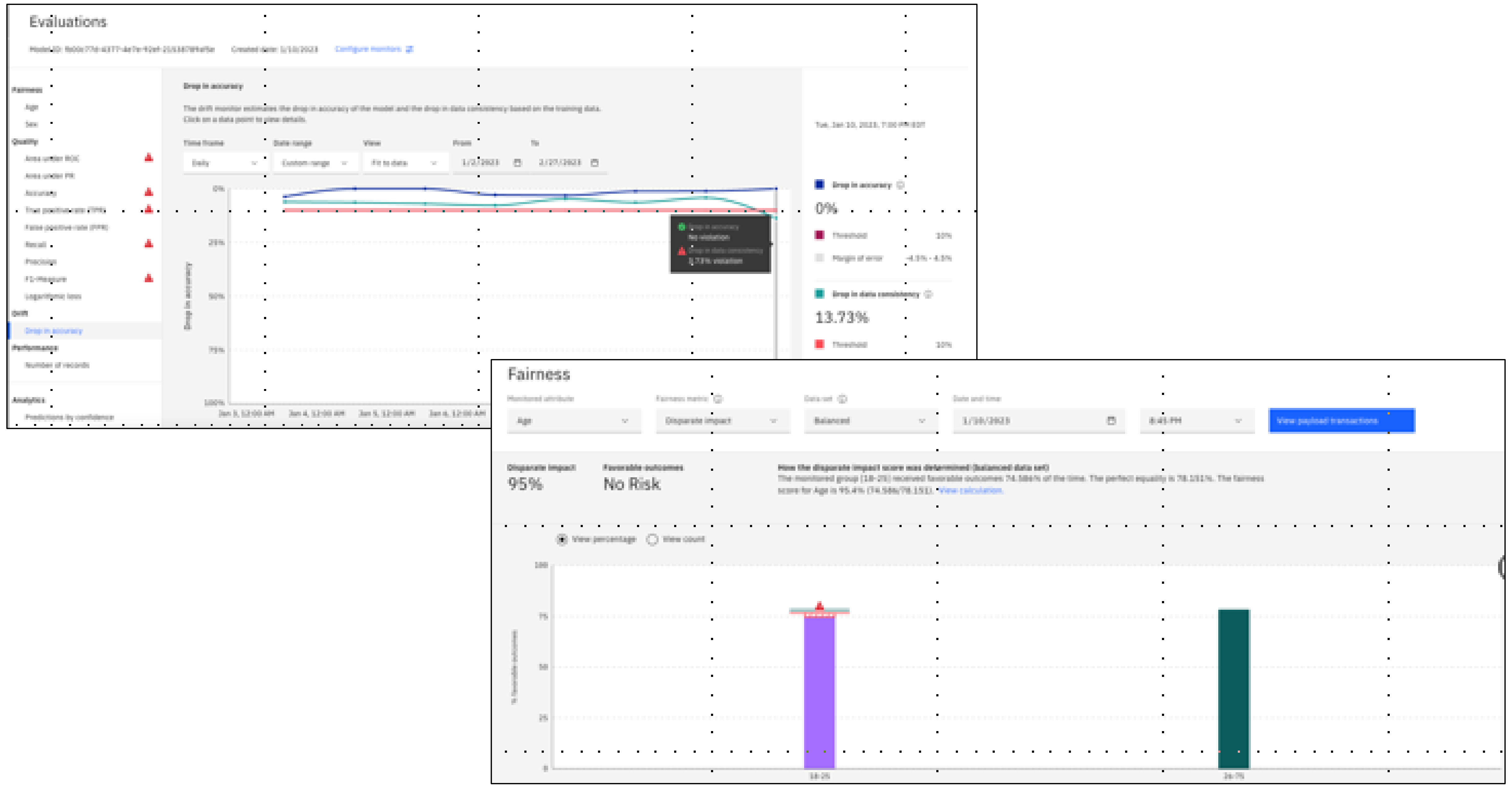
Quality

Area under ROC	Area under PR	Accuracy	True positive rate (TPR)	False positive rate (FPR)	Recall	Precision	F1-Measure	Logarithmic loss
0.76 0.19 lower threshold violation	0.76	0.82	0.57	0.05	0.57	0.87	0.69	0.43

		Prediction		
		No Risk	Risk	Total
Actual	No Risk	62	3	65
	Risk	15	20	35
	Total	77	23	100

Drive model quality

- Monitor model quality metrics for accuracy, precision, recall, and more.
- Receive alerts when the value goes beyond configurable thresholds
- Define custom metrics to track quality for model predictions



Evaluations

Drop in accuracy

The drift monitor estimates the drop in accuracy of the model and the drop in data consistency based on the training data. Click on a data point to view details.

Drop in accuracy: 0%

Drop in data consistency: 13.73%

Fairness

Disparate impact: 95%

Favorable outcomes: No Risk

How the disparate impact score was determined (balanced data set)
The monitored group (18-25) received favorable outcomes 74.586% of the time. The perfect equality is 78.151%. The fairness score for Age is 95.4% (74.586/78.151). View calculation.

Detect and mitigate bias and drift

- Automate the detection of bias and drift and associated datapoints
- Detect biases in runtime, identify impacts automatically to comply with regulations
- Identify drift in accuracy with predicted model performance
- Set alerts in the risk management dashboard
- Provide metrics and data to help data scientist troubleshoot bias

watsonx.governance

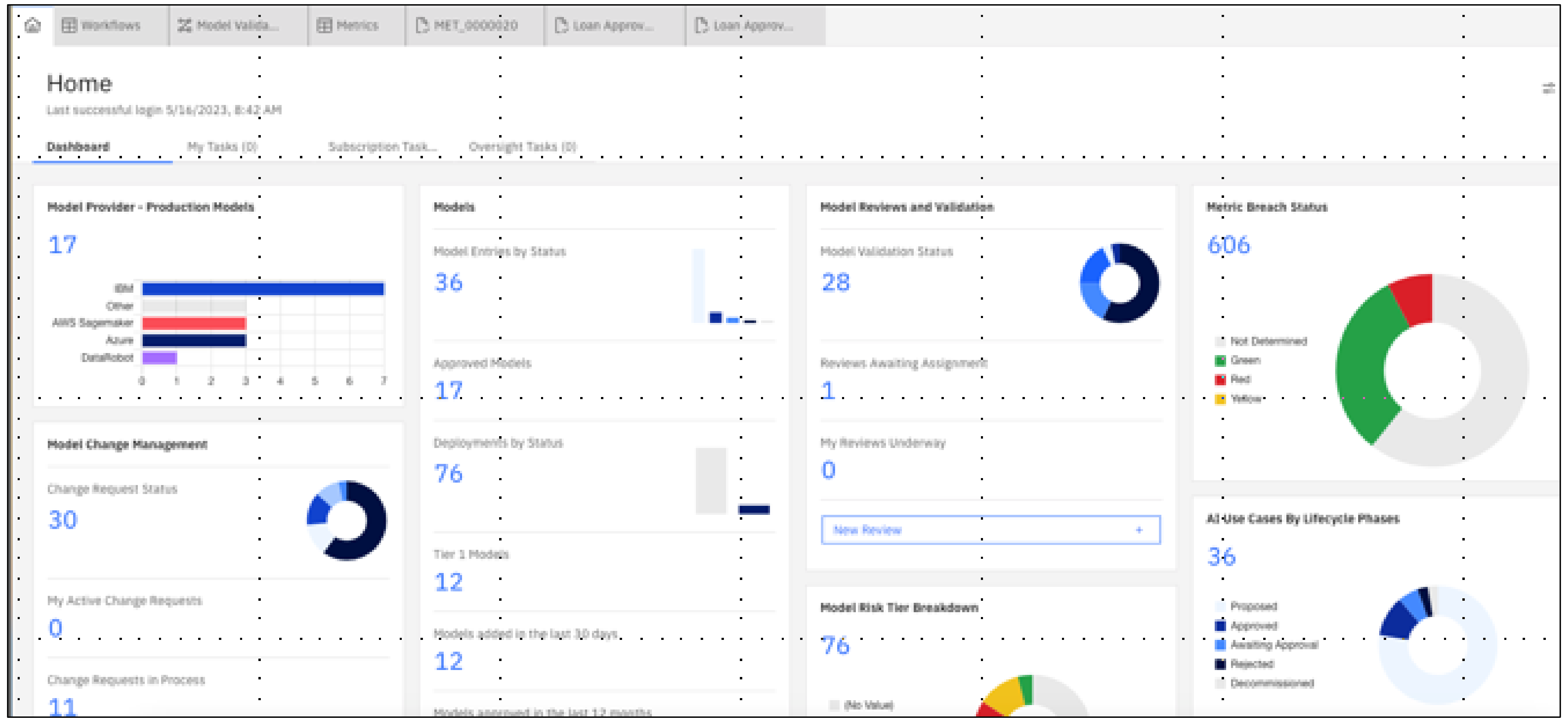
Manage risk across the enterprise

Implement model risk management

- Automate the monitoring of production models to match pre-production settings
- Enable customized tests to compare model performance
- Generate outcome reports
- Integrate with AI factsheets for model documentation

Drive transparent model results

- Pull together models from multiple platforms
- View development status, model performance and alerts for emerging issues
- Monitor and trigger workflows for model validation, retraining and performance issues



Interactive dashboard summarizes entire AI landscape

watsonx.governance

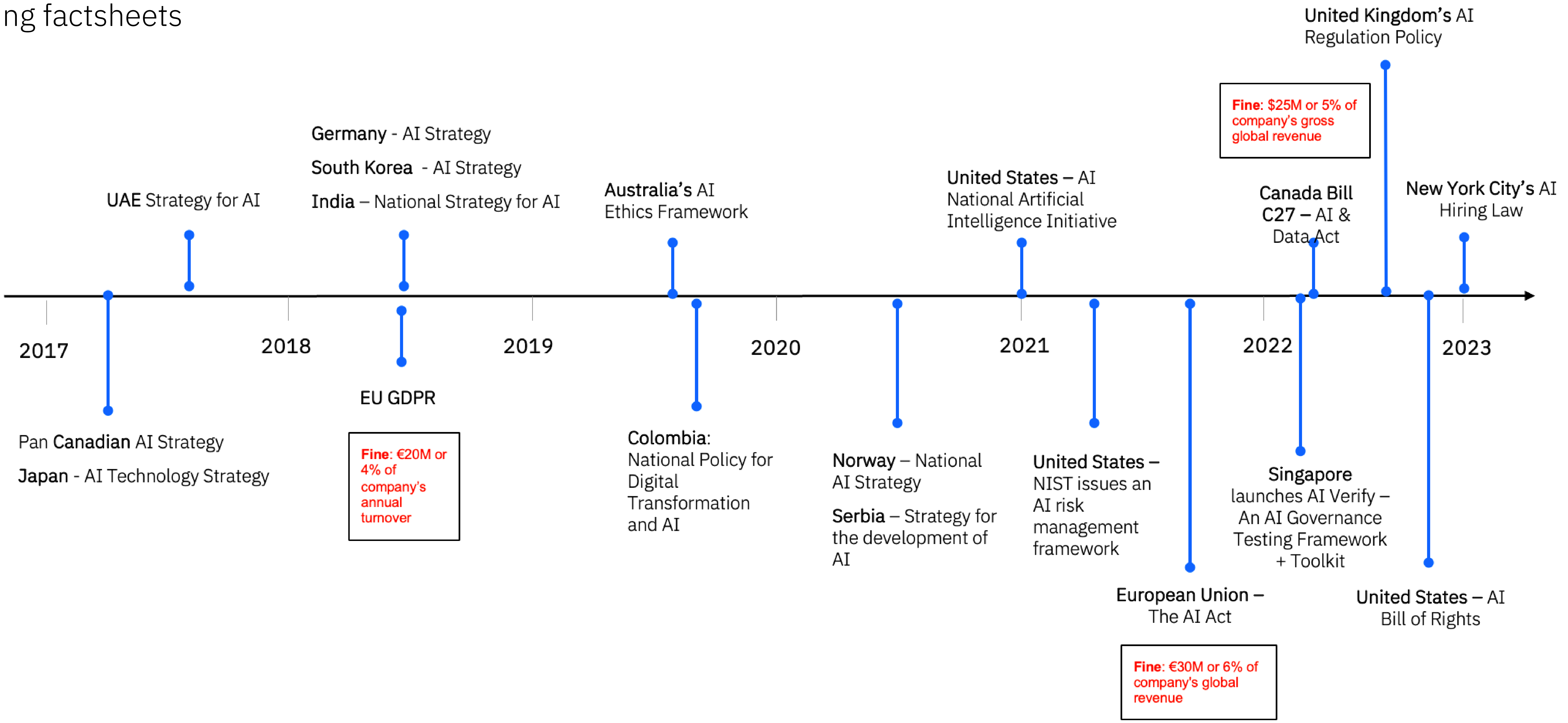
Adhere to regulatory compliance

Drive AI model explainability

- Automate the translation of AI regulations into enforceable standards and policies
- Track provenance, document model performance against KPIs
- Use dynamic dashboards and automated collaborative tools
- Document model facts using factsheets

Better meet growing AI and industry regulations

- Avoid costly fines and audits due to noncompliance and quickly respond to regulatory change
- Efficiently process large volumes of regulations and industry standards
- Ensure stakeholders need to be informed on regulatory compliance.



2017

Pan Canadian AI Strategy

Japan - AI Technology Strategy

UAE Strategy for AI

2018

Germany - AI Strategy

South Korea - AI Strategy

India – National Strategy for AI

EU GDPR

Fine: €20M or 4% of company's annual turnover

2019

Australia's AI Ethics Framework

Colombia: National Policy for Digital Transformation and AI

2020

Norway – National AI Strategy

Serbia – Strategy for the development of AI

United States – AI National Artificial Intelligence Initiative

2021

United States – NIST issues an AI risk management framework

European Union – The AI Act

Fine: €30M or 6% of company's global revenue

2022

Canada Bill C27 – AI & Data Act

Fine: $25M or 5% of company's gross global revenue

Singapore launches AI Verify – An AI Governance Testing Framework + Toolkit

United Kingdom's AI Regulation Policy

United States – AI Bill of Rights

2023

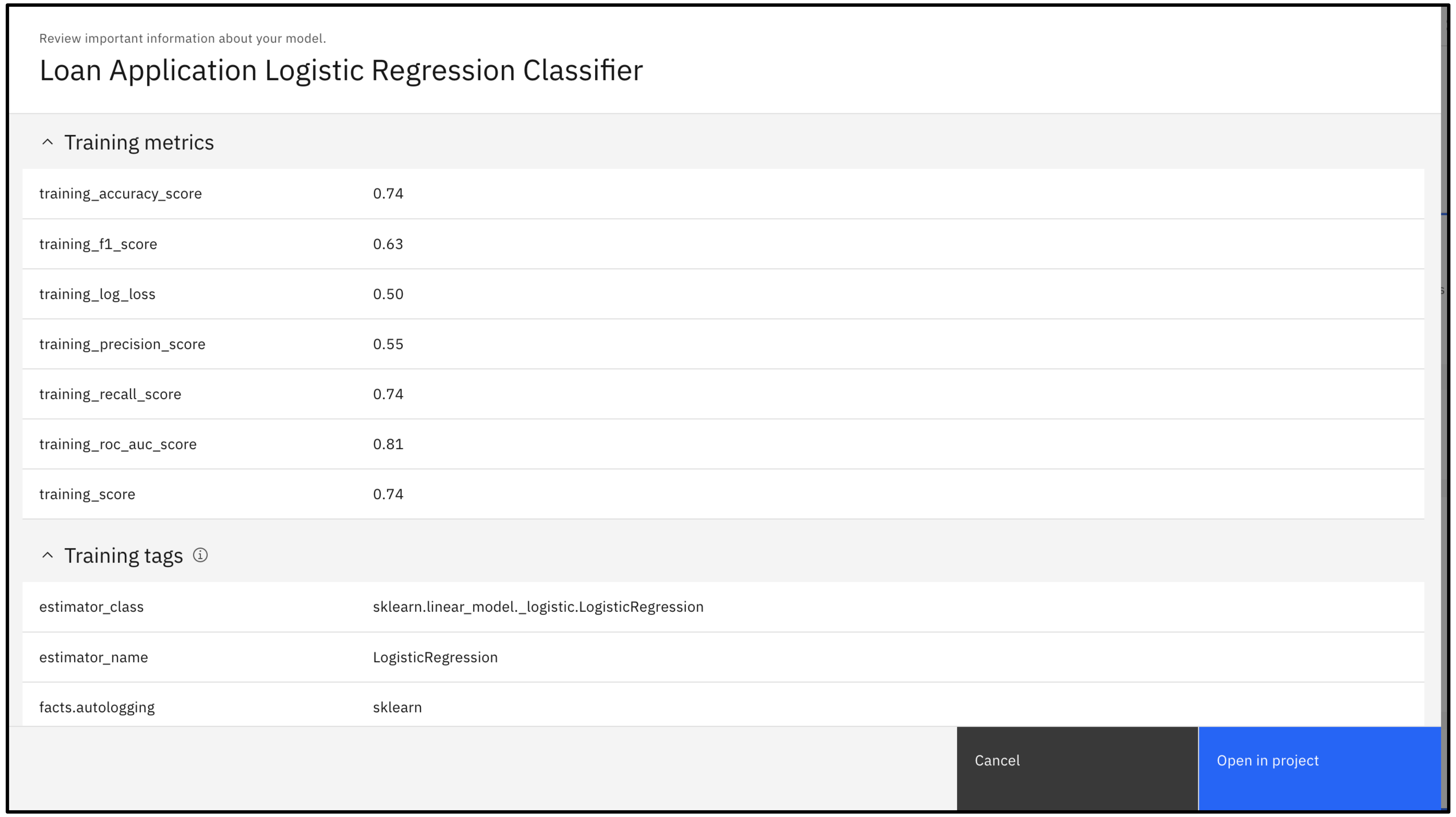
New York City's AI Hiring Law

watsonx.governance

Adhere to regulatory compliance

Model auditability with factsheets

- Automate documentation of key AI metadata, explain transaction level decisions in model runtime
- Provide a singular view of facts across the model lifecycle
- Facilitate subsequent enterprise validation, understand how the model will behave in different business situations
- Support audits, and requests for model facts from auditors, management, stakeholders and customers



Review important information about your model.

Loan Application Logistic Regression Classifier

^ Training metrics

training_accuracy_score	0.74
training_f1_score	0.63
training_log_loss	0.50
training_precision_score	0.55
training_recall_score	0.74
training_roc_auc_score	0.81
training_score	0.74

^ Training tags ⓘ

estimator_class	sklearn.linear_model._logistic.LogisticRegression
estimator_name	LogisticRegression
facts.autologging	sklearn

Why IBM?

Why IBM?

Open

IBM’s AI is based on the best open technologies available

Trusted

IBM’s AI is transparent, responsible, and governed

Targeted

IBM’s AI is designed for enterprise and targeted at business domains

Empowering

IBM’s AI is for value creators, not just users

IBM’s differentiators

Invested billions in innovation

→ State of the art capabilities from a rich pipeline of IBM Research innovations

→ Enhanced versions of the AI360 toolkits comprising 150+ algorithms and metrics

→ Technology is used internally “at IBM scale” with IBM’s AskHR bot

→ Very active in the technical and regulatory communities

Most comprehensive offering

→ Only vendor to address all three pillars of AI governance: lifecycle governance, risk management, and regulatory compliance

→ Deep expertise in enterprise GRC as well as enterprise AI

→ Not just a governance platform, also integrates with AI platforms

→ Enhance people, process, and technology, informed by IBM’s own internal efforts

Designed to govern AI in any platform

→ Consistent governance regardless of where the AI is created or deployed

→ Detect unfair bias and perform explanations at design time on any platform using Python notebooks

→ Capture model metadata from any platform

→ Monitor models (on any platform) for model health, accuracy, drift, and bias

Global leadership & collaboration

Our principles and pillars in practice / Partnerships

U.S. National AI Advisory Committee (NAIAC)

Chief Privacy Officer Christina Montgomery named to NAIAC and U.S. Chamber of Commerce Commission on Competition, Inclusion and Innovation

European Commission Expert Group on AI

Defined the ethics guidelines for trustworthy AI

Partnership on AI

Brings together diverse global voices to define best practices for beneficial AI

IBM is a founding member

IEEE Global Initiative on AI Ethics

Supports development of AI that prioritizes ethical considerations

World Economic Forum's Global AI Action Alliance

Guides the responsible development of AI

Co-chaired by Arvind Krishna, IBM Chairman and CEO

ITU AI for Good Global Summit

Global and inclusive United Nations platform on using AI to achieve the UN Sustainable Development Goals

MIT-IBM Watson AI Lab

Research focused on healthcare, security and finance using the IBM Cloud, AI platform, blockchain and quantum computing

Data & Trust Alliance

Develops new practices and tools to advance the responsible use of data and AI across industries and disciplines

IBM customers are maturing their AI governance

More than 80% say that they'll commit 10% or more of their total AI budget to meeting regulatory requirements by 2024 and 45% are planning to spend at least 20%.

Accenture - From AI compliance to competitive advantage, 2022

Prepare for audit and regulatory compliance

North American Bank, multiple data science stacks, 1000s of models.

Manual audit process taking months of work.

Invested in IBM software for its completeness and ability to work with existing technology.

Proactively mitigate bias in the hiring process

North American retailer wanted to meet its commitments as a fair employer.

Invested in IBM software to monitor and actively look for potential bias in their hiring systems.

us open

A novel way to look at AI bias

AI-assisted curation of match highlights, available 2 minutes after the match ends.

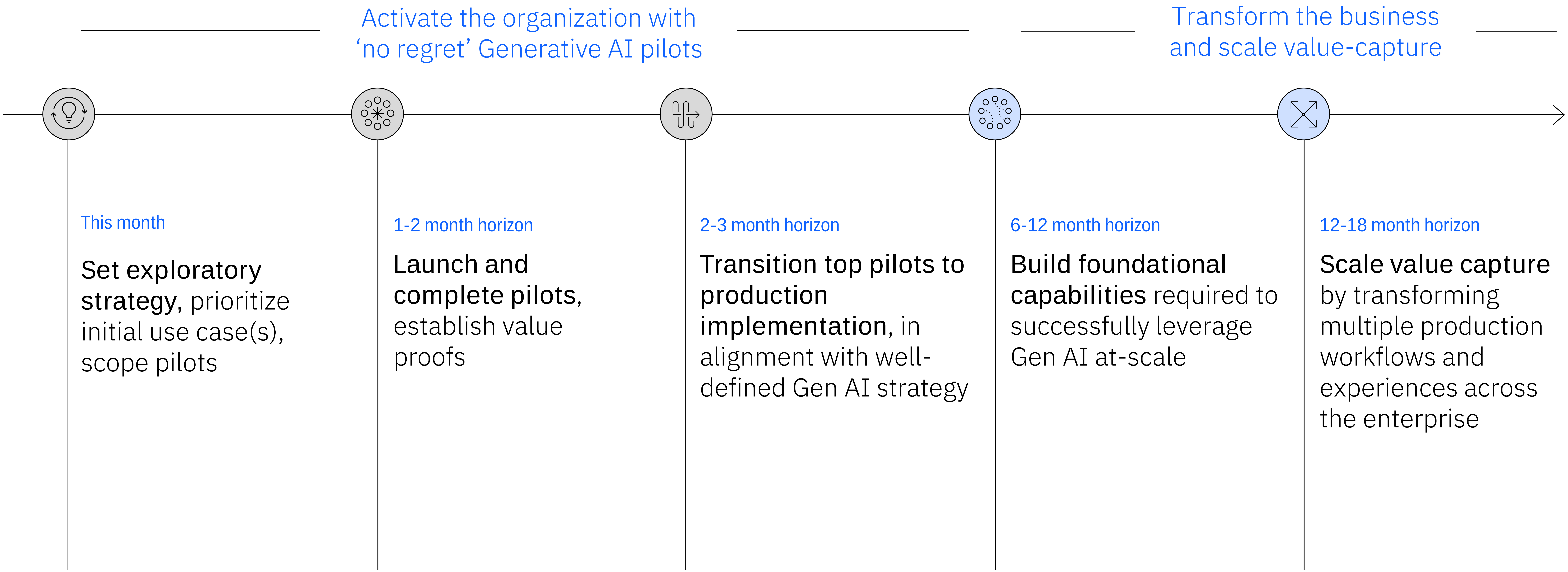
Excitement score is biased by player rank and the court where the match is played.

Post-processing de-biasing applied to increase court fairness from 71% to 82% without impacting overall accuracy.

Getting started

IBM can help you along your AI transformation journey to unlock value at-scale

From rapid activation in a “no regrets” pilot to a holistic transformation effort



Activate the organization with ‘no regret’ Generative AI pilots

Transform the business and scale value-capture

This month

Set exploratory strategy, prioritize initial use case(s), scope pilots

1-2 month horizon

Launch and complete pilots, establish value proofs

2-3 month horizon

Transition top pilots to production implementation, in alignment with well-defined Gen AI strategy

6-12 month horizon

Build foundational capabilities required to successfully leverage Gen AI at-scale

12-18 month horizon

Scale value capture by transforming multiple production workflows and experiences across the enterprise

Three ways to get started with **watsonx.governance** today

IBM’s investment in partnering with you

REQUEST A DEMO

Experience **watsonx.governance** and see core capabilities with a free demo

Available now

CLIENT BRIEFING

Discussion and custom demonstration of IBM’s generative AI **watsonx** point-of-view and capabilities. Understand how **watsonx.governance** can be leveraged in your AI strategy.

2-4 hours

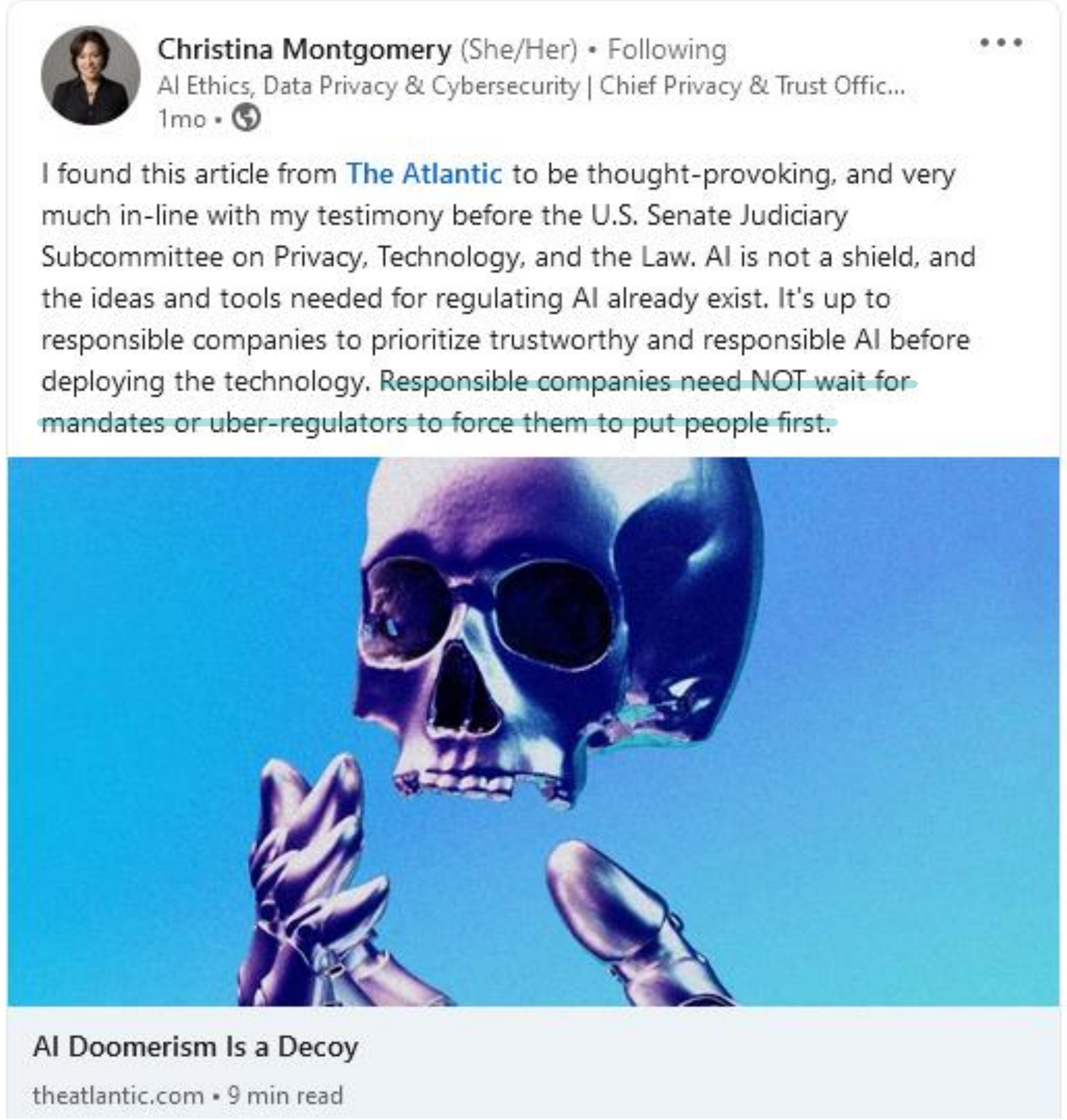
Onsite or virtual

PILOT PROGRAM

watsonx.governance pilot developed with IBM AI engineers. Prove **watsonx.governance** value for the selected use case(s) with a plan for adoption.

1-4 weeks

No reason to wait



Christina Montgomery (She/Her) • Following

AI Ethics, Data Privacy & Cybersecurity | Chief Privacy & Trust Offic...

1mo •

I found this article from The Atlantic to be thought-provoking, and very much in-line with my testimony before the U.S. Senate Judiciary Subcommittee on Privacy, Technology, and the Law. AI is not a shield, and the ideas and tools needed for regulating AI already exist. It's up to responsible companies to prioritize trustworthy and responsible AI before deploying the technology. Responsible companies need NOT wait for mandates or uber-regulators to force them to put people first.

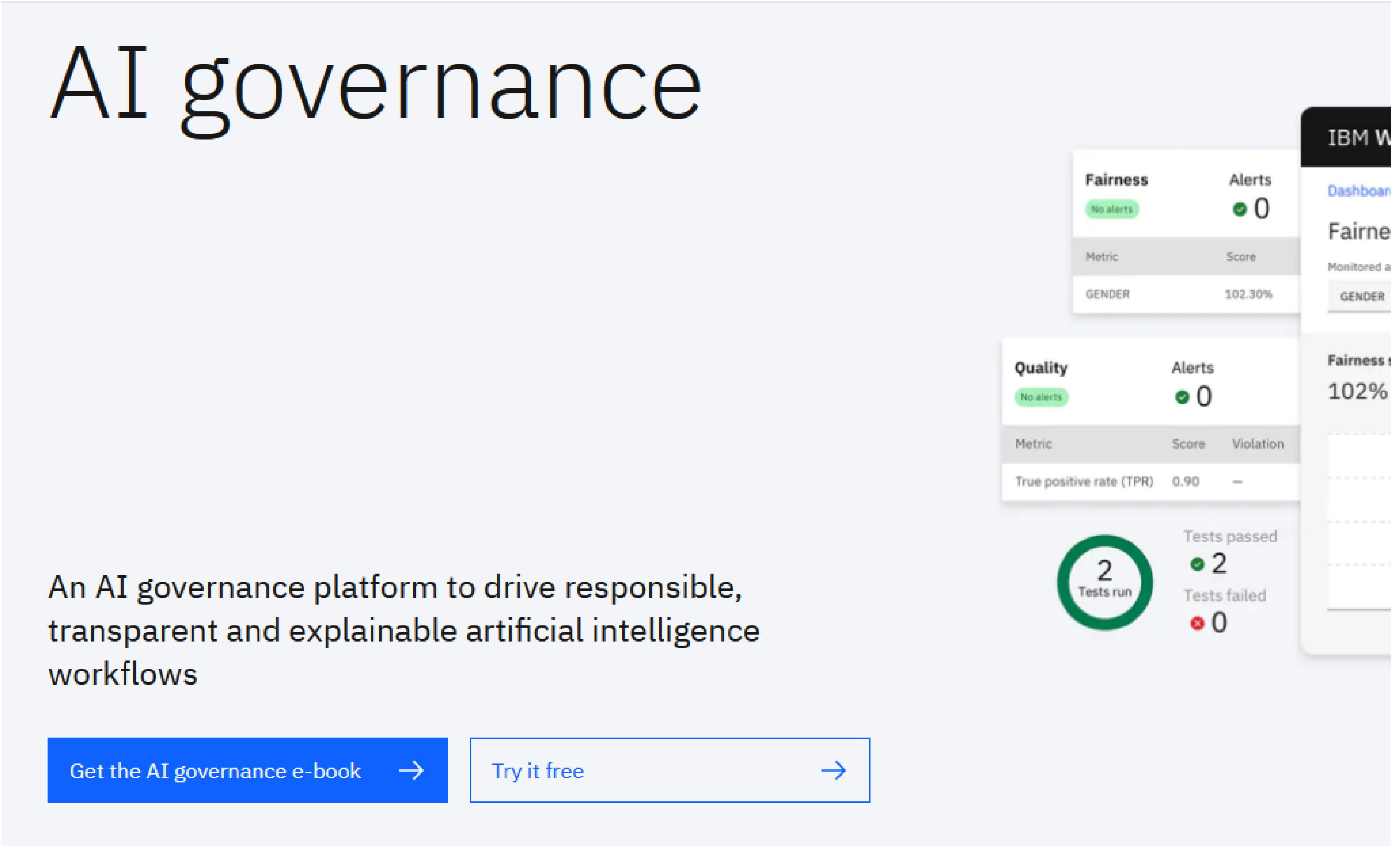
AI Doomerism Is a Decoy

theatlantic.com • 9 min read

A client after seeing a demonstration of IBM AI Governance:

“The whole organization is rushing into generative AI, but we don’t even have this in place yet for our existing models.”

Available now, even better in the future:



AI governance

An AI governance platform to drive responsible, transparent and explainable artificial intelligence workflows

Get the AI governance e-book →

Try it free →

Fairness
No alerts
Alerts 0

Metric	Score
GENDER	102.30%

Quality
No alerts
Alerts 0

Metric	Score	Violation
True positive rate (TPR)	0.90	—

2 Tests run

Tests passed 2

Tests failed 0

IBM