Operationalizing AI

Hans-Petter Dalen
Business Executive for EMEA

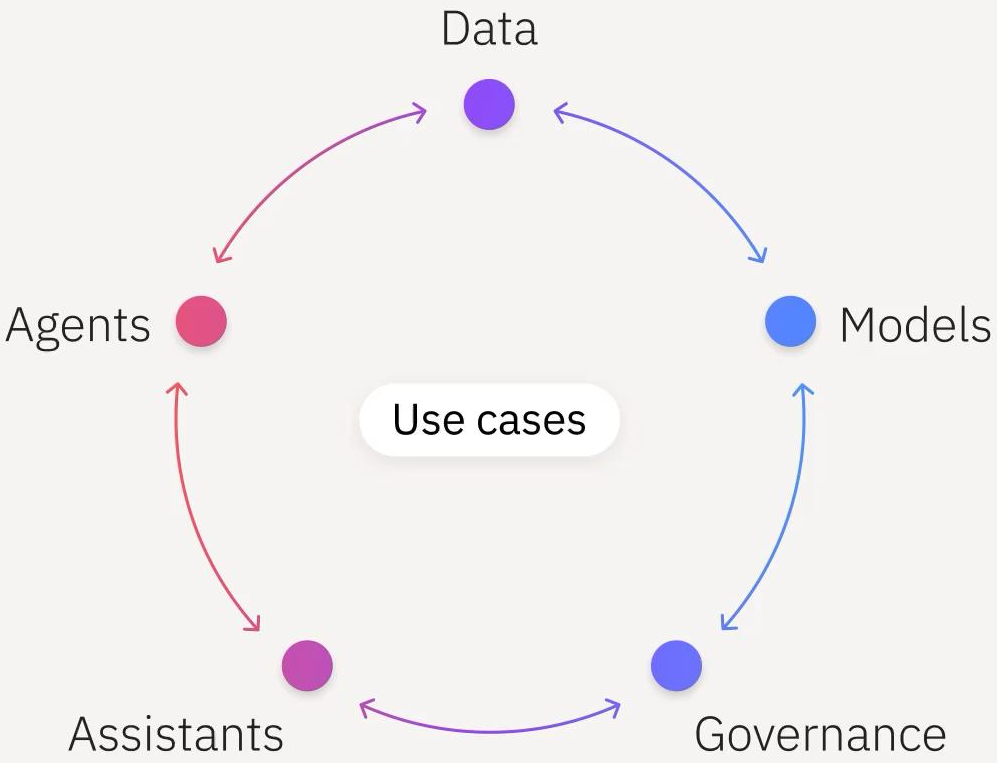
AI for Business

watsonx.

Business impact of AI

"AI could enable automation of up to 70% of business activities"
(McKinsey)

AI building blocks to the future



IBM is “client zero” for AI governance at scale

People & Process

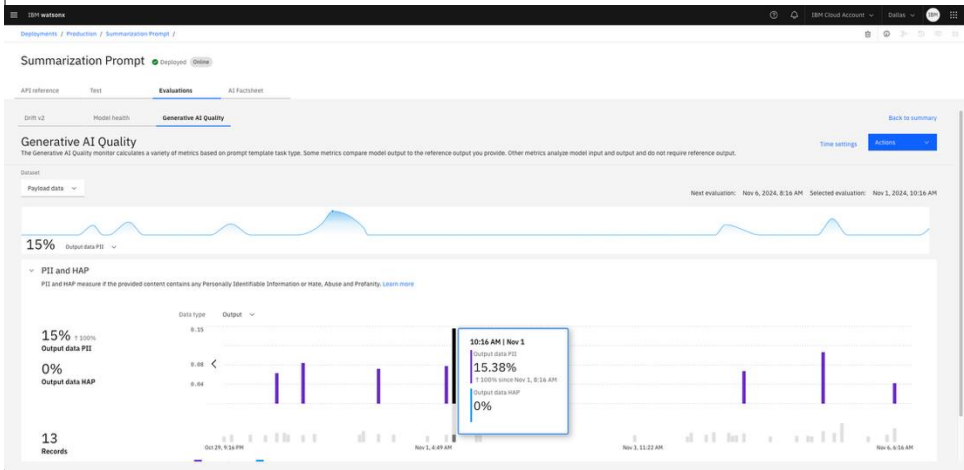
- Accountability model
- Centralized governance and decision making (AI Ethics Board)
- Focals in the business units
- Policies and Principles
- Ethics by Design
- Awareness and education
- Integrated Governance Program

Technology

- watsonx.governance
- Integration with systems of record
- Self-service data gathering
- Standardized assessments
- Use case review process
- Extensive dashboards
- AI monitoring

Results

- **>6,500** applications and processes managed
- **6 weeks** time for a recent new enterprise-wide compliance campaign
- **Days to minutes** projected reduction in time to evaluate AI solutions during build stage
- **2x** increase in Gen AI metrics evaluated during solution validation
- **14** health, drift, safety metrics continuously monitored in production
- Foundation for quick adoption of new technical advances



Your principles of trust are the foundation of all your ethical decisions, including use case reviews

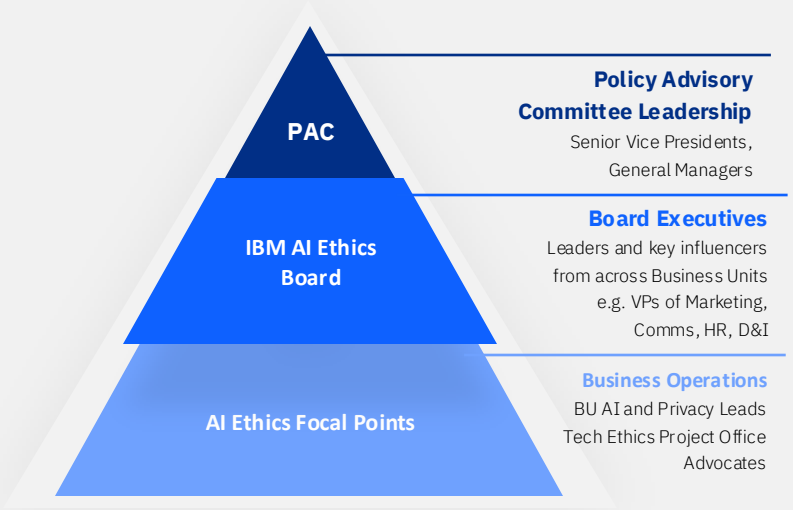
IBM Principles for Trust and Transparency

1. The purpose of AI is to augment — not replace — human intelligence
2. Data and insights belong to their creator
3. New technology, including AI systems, must be transparent and explainable

IBM Pillars of Trust

- Explainability
- Transparency
- Fairness
- Privacy
- Robustness

Centralized organizational AI governance enables top-down and bottom-up engagement



Benefits of a centralized AI governance structure:

- Defined accountability and ownership starting at the top, with leaders sponsoring a culture focused on responsible AI
- Consistent governance and decision-making as the enterprise develops, deploys, and uses AI and other technologies
- A robust and scalable framework for AI governance and ethics

IBM as Client Zero

Training available for all roles fosters a culture of responsible AI within their day-to-day jobs

All	Business Leaders	Product/Offering Managers	Data Scientists	Developers	Sales	Designers
Annual Cybersecurity and Data Privacy Education*	AI Associate Leader Badge	AI Associate for Offering and Product Managers Badge	AI Associate for Data Scientists Badge	Ethics by Design Learning Plan	AI Associate Leader Badge	Everyday Ethics for AI
Business Unit based Generative AI Training*	Tech Ethics Use Case Review	Completing the Tech Ethics Use Case Review	Ethics by Design Learning Plan	Everyday Ethics for AI	Business Partner Integrity – AI Ethics module+	Foundation Models for Designers Guide
Trustworthy AI and AI Ethics Foundations Badge	AI for Leaders	Completing the Algorithmic Impact Assessment	Trustworthy AI for Practitioners	Trustworthy AI for Practitioners	Integrity in Supplier Relationships – AI Ethics module+	Trustworthy AI for Practitioners
AI Ethics Learning Plan		Trustworthy AI for Practitioners	Assess & Mitigate Risk	Regional TAI courses		Regional TAI courses
Ethics by Design Learning Plan			Regional TAI courses	Gen AI COE Monthly Webinars		
Academy of Technology Initiatives				GEN AI Red Team		
Trustworthy AI COE Monthly Webinars						

* Required

\+ Vendors & Business Partners

AI Governance and AI Security

The Sandwich Dilemma

Undesired Results

Desired Results

AI Governance

AI Security

Human

AI

https://food.ec.europa.eu/horizontal-topics/farm-fork-strategy_en

The reasons we MUST have Trustworthy AI … The 4 Rs

WHY

Regulatory Fines

EU Artificial Intelligence Act

Noncompliance case	Proposed fine
Breach of AI Act prohibitions	Fines up to €35 million or 7% of total worldwide annual turnover (revenue), whichever is higher
Noncompliance with the obligations set out for providers of high-risk AI systems or GPAI models, authorized representatives, importers, distributors, users or notified bodies	Fines up to €15 million or 3% of total worldwide annual turnover (revenue), whichever is higher
Supply of incorrect or misleading information to the notified bodies or national competent authorities in reply to a request	Fines up to €7.5 million or 1.5% of total worldwide annual turnover (revenue), whichever is higher

https://artificialintelligenceact.eu/

The reasons we MUST have Trustworthy AI … The 4 Rs

Regulatory Fines

Reputational Damage

The customer then posted the chat, which had gone viral with 1.3 million views and over 20 thousand likes.

The reasons we MUST have Trustworthy AI … The 4 Rs

WHY

Regulatory Fines

Reputational Damage

Revenue Loss
(primary / secondary)

The reasons we MUST have Trustworthy AI … The 4 Rs

WHY

Regulatory Fines

Reputational Damage

Revenue Loss
(primary / secondary)

Running Operations Impacts


Gartner

Who We Serve Our Solutions Latest Insight Conferences Webinars AI

Newsroom Topics Media Contacts Media Resources Insights Archive

Newsroom / Information Technology / Press Release

Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027

SYDNEY, Australia, June 25, 2025

Analysts to Explore Agentic AI Trends During Gartner IT Symposium/Xpo, September 8-10 on the Gold Coast

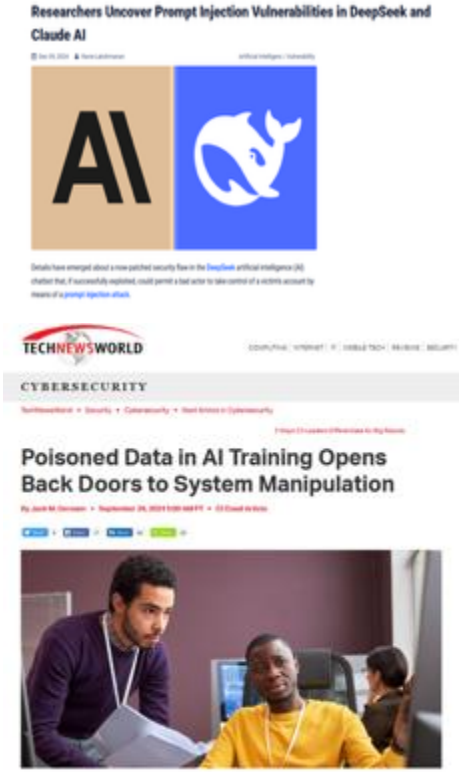
due to:-

- escalating costs
- unclear business value
- inadequate risk controls

https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

High-profile AI security breaches are already happening …

In the news



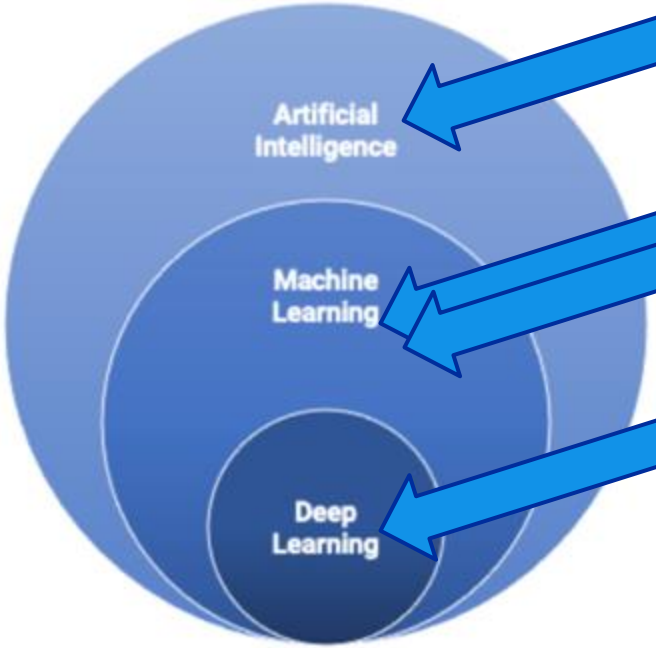
… and it will get worse

*"By 2027, **more than 40% of AI-related data breaches will be caused by the improper use of generative AI** (GenAI) across borders, according to Gartner, Inc."*

Source: Gartner (Feb 2025)

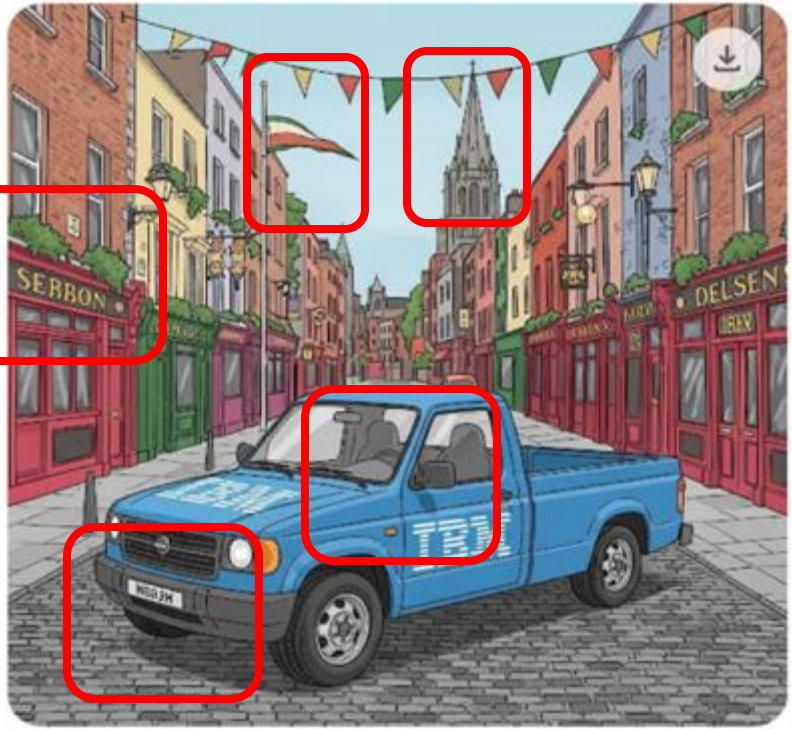
Evolution of AI and its vulnerabilities

HOW



Here's a drawing of an IBM pickup truck in Dublin:

Do you need any other images?



How AI can misbehave

(In general these are the attributes when AI misbehaves)

AI Governance (in->out)

Passive Anomalies
(accidental)

- Quality Assu Miss
- Clumsy & Lo
- No Malic nt
- Rem eeds no escalation
- F cause Analysis
- nable Impacts

Active Anomalies
(coerced)

- Targeted n abuse
- Focused
- Malic nt
- R on needs escalation due to curity implications
- nsive Root cause Analysis needed

AI Security (in<-out)

Trustworthy AI

- Financial gain
- Revenge
- Ideological beliefs
- Coercion
- Ignorance

AI can be considered you new INSIDER THREAT

HOW

What organizations need from AI Governance

Lifecycle governance
Managing AI assets across the entire lifecycle

Evaluate, Monitor, Detect, and Protect
Evaluate and monitor for fairness, accuracy, and drift, while protecting from harmful content

Risk Management & Compliance
Way to manage AI activities and risks from initial requests, production deployments, to compliance

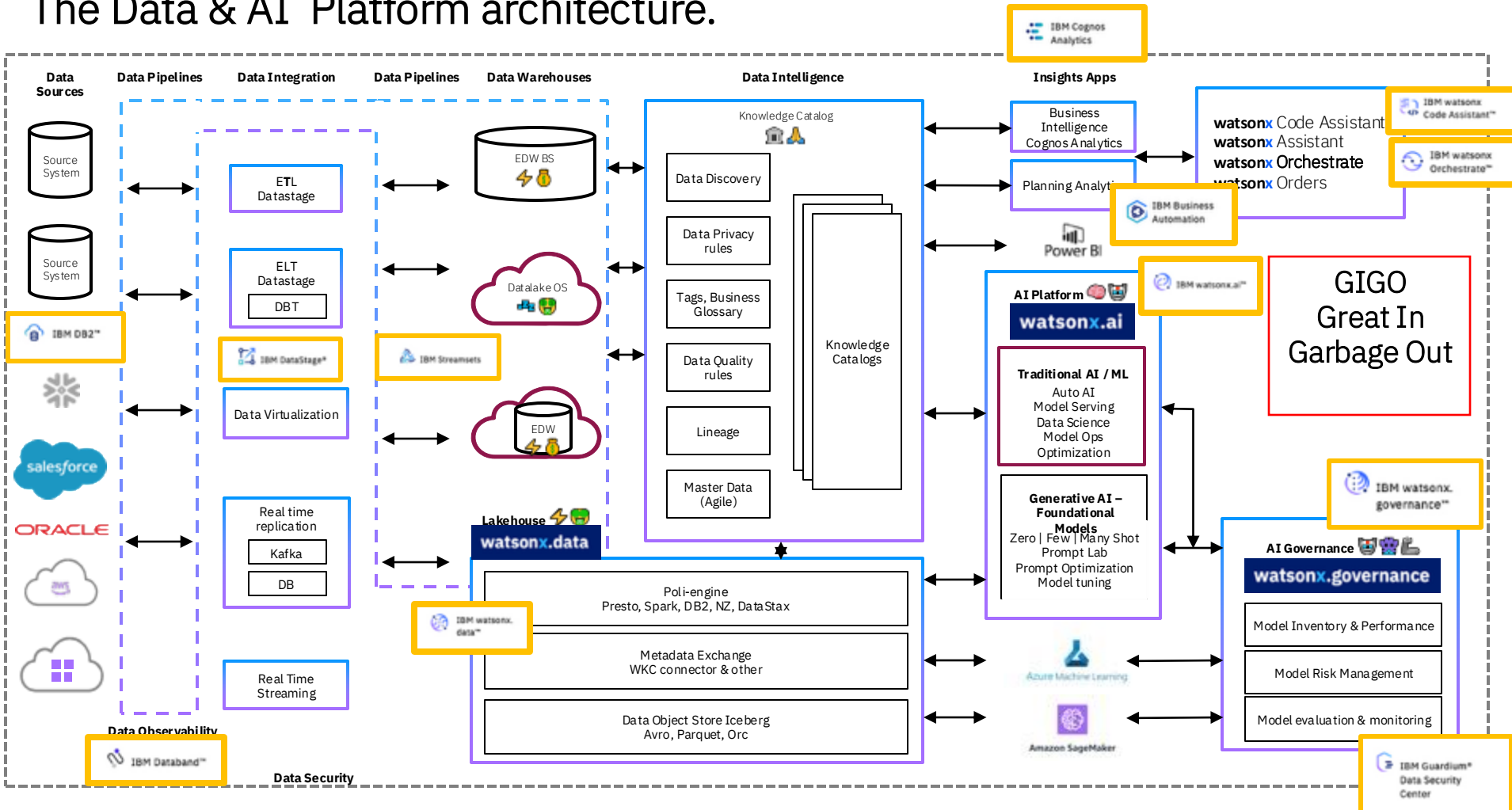
What organizations need from AI Security

Visibility
Understand where AI assets and services are deployed and in use within the organization

Security for Data
Monitor data sources, especially sensitive and training data, leveraged in AI deployments to prevent breaches

Security for Models & Applications
Apply security controls to mitigate vulnerabilities and respond to threats quickly

The Data & AI Platform architecture.



Understanding Risk

Evolving elements of AI risk

AI agent risks	**Amplified** • Misaligned actions • Discriminatory actions • Over- or under-reliance • Unauthorized use • Exploit trust mismatch • Sharing IP/PI/confidential information with user • Unexplainable or untraceable actions	**New** • Data bias • Redundant actions • Attack on an AI agent's external resources • Function-calling hallucination • Sharing IP/PI/confidential information with tools
+		
Foundation model risks	**Amplified** • Data bias • Data curation • Data acquisition • Data usage rights • Data transparency • Data provenance • Accountability • Source attribution • Impact on jobs • Data privacy rights • Evasion attack • Nonconsensual use • Improper usage • Over/under reliance • Unexplainable output • Human exploitation • Impact on environment • Impact on human agency	**New** • Downstream-based retraining • PII in prompt • IP in prompt • Confidential data in prompt • Prompt-based attack • Output bias • Copyright infringement • Toxic output • Dangerous advice • Legal accountability • Generated content ownership • Generated content IP • Spreading disinformation • Toxicity • Dangerous use • Non-disclosure • Impact on cultural diversity • Harmful code generation • Exposing personal information • Hallucination • Traceability • Impact on education
+		
Traditional AI risks	• Data usage restrictions • Data transfer restrictions • Personal information in data • Reidentification • Unrepresentative data • Data poisoning	• Decision bias • Model usage rights restrictions • Lack of system transparency • Lack of model transparency • Impact on affected communities

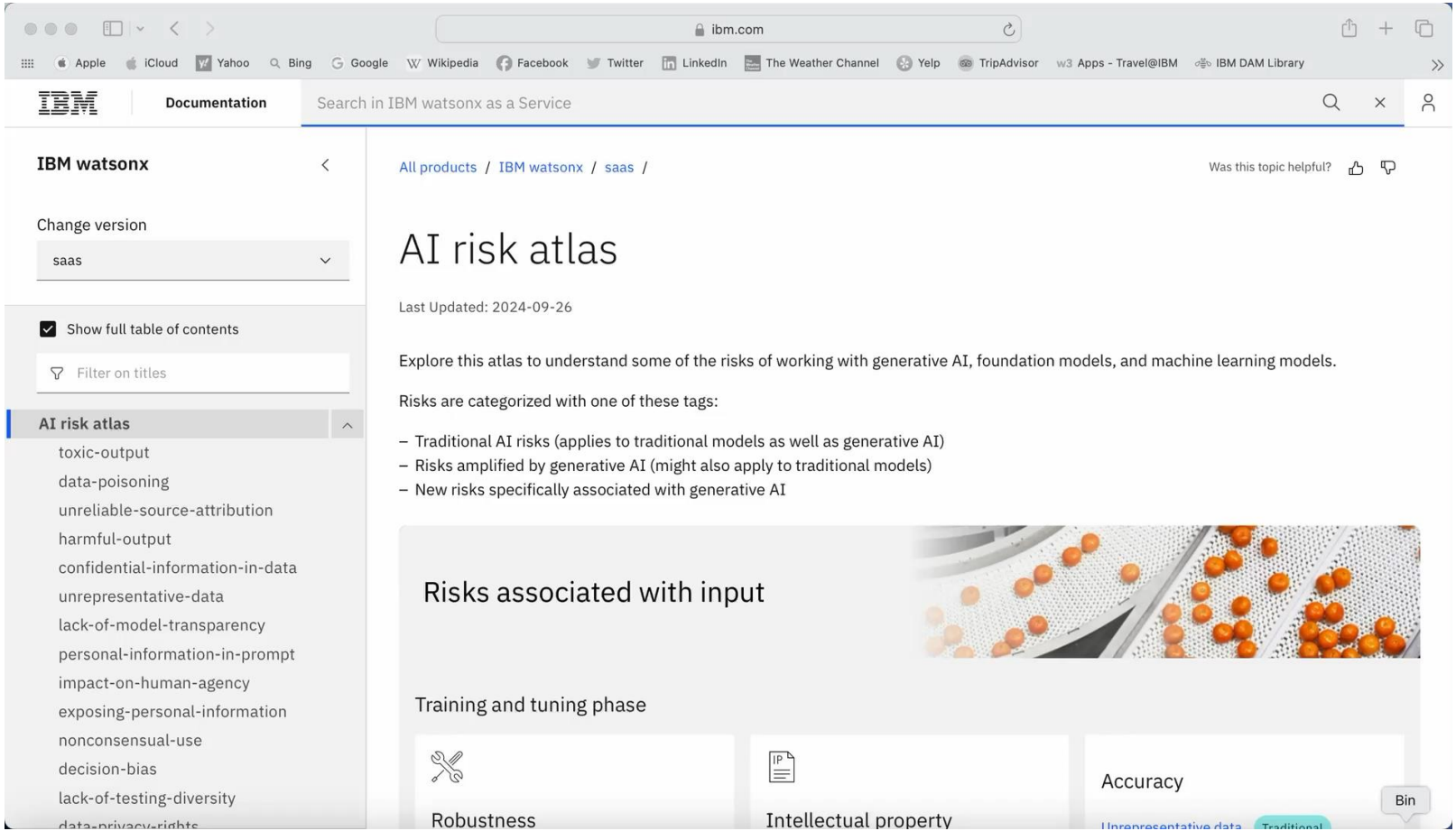
Regulatory Risk

Reputational Risk

Operational Risk

IBM AI Ethics Board - Opportunities, risks and mitigations – Foundation models
IBM AI Ethics Board - Opportunities, risks and mitigations – AI agents
IBM AI Risk Atlas

The IBM AI Risk Atlas



Agentic AI

Risks and Challenges

New

Emerging areas *intrinsic to* agentic AI

Risks

- Unsupervised autonomy
- Data bias
- Redundant actions
- Attack on AI agent's external resource
- Tool choice hallucination
- Sharing IP/PI/confidential information

Challenges

- Reproducibility
- Traceability
- Attack surface expansion
- Harmful and irreversible consequences

Amplified

Known areas *intensified by* agentic AI

Risks

- Misaligned actions
- Discriminatory actions
- Over- or under-reliance
- Unauthorized use
- Exploit trust mismatch
- Unexplainable or untraceable actions
- Lack of transparency

Challenges

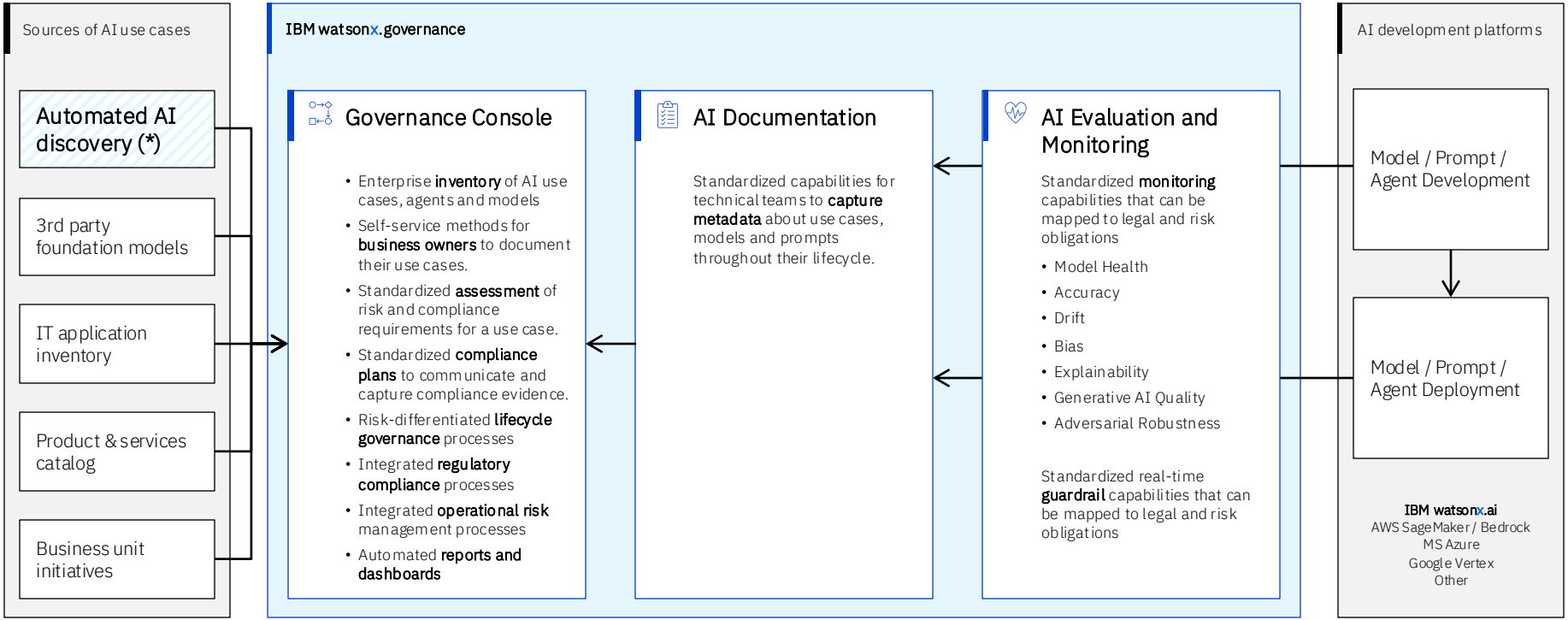
- Evaluation
- Accountability
- Compliance
- Mitigation and maintenance
- Infinite feedback loops
- Shared model pitfalls

For more information
on IBM’s perspective

Read *AI agents: Opportunities, risks, and mitigations*, a deep-dive into the unique risks posed by AI agents and potential mitigations, written by the IBM AI Ethics Board.

Scan the QR code to
access the paper:

watsonx.governance is software to scale governance of AI



(*) Through integration with IBM Guardium AI Security