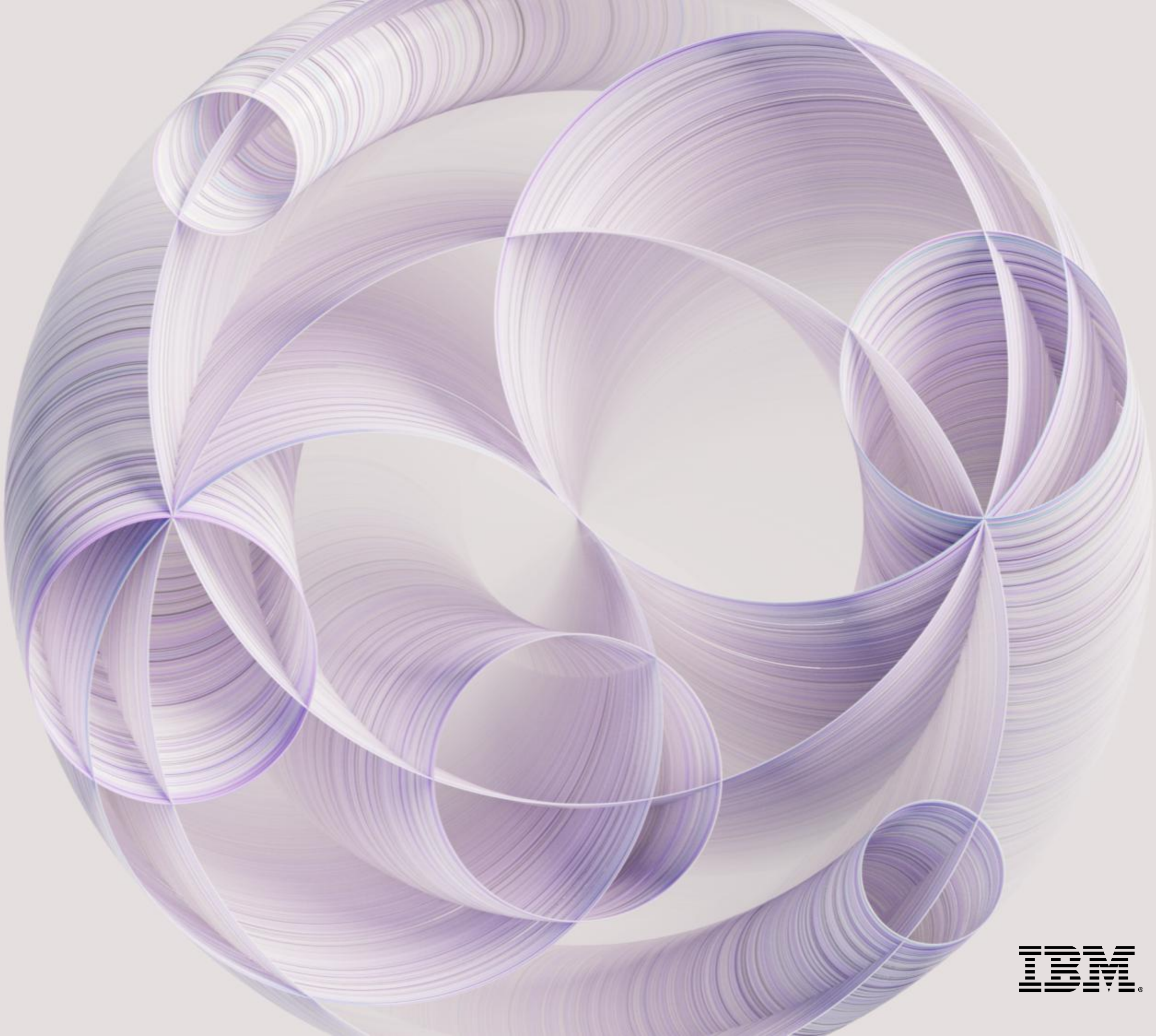
Agentic AI :

Understanding the risks

Hans-Petter Dalen
Business Executive for EMEA

AI for Business

watsonx.



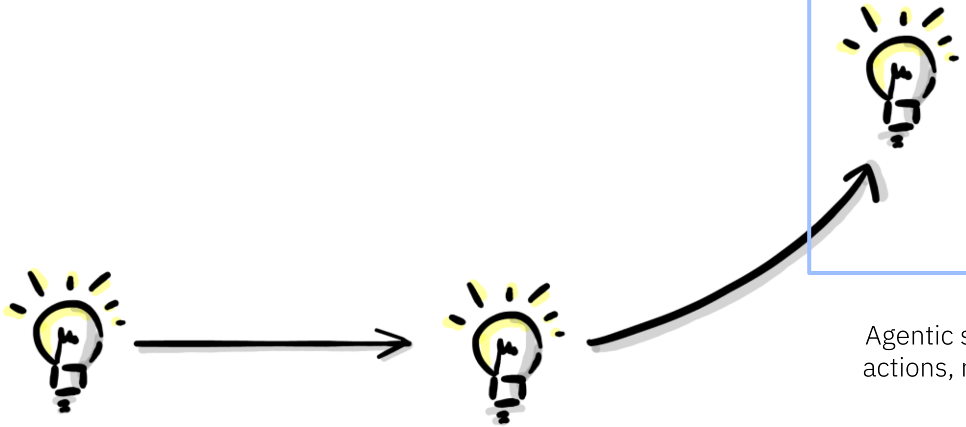
IBM

Sell me this pen.
It's AI powered.
EVERY COMPANY
RIGHT NOW

Business impact of AI

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

(McKinsey)

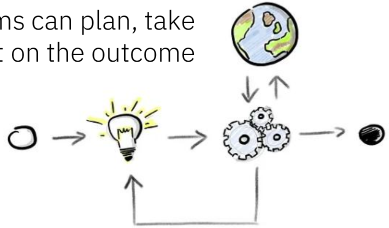


2022
AI can **predict** for you

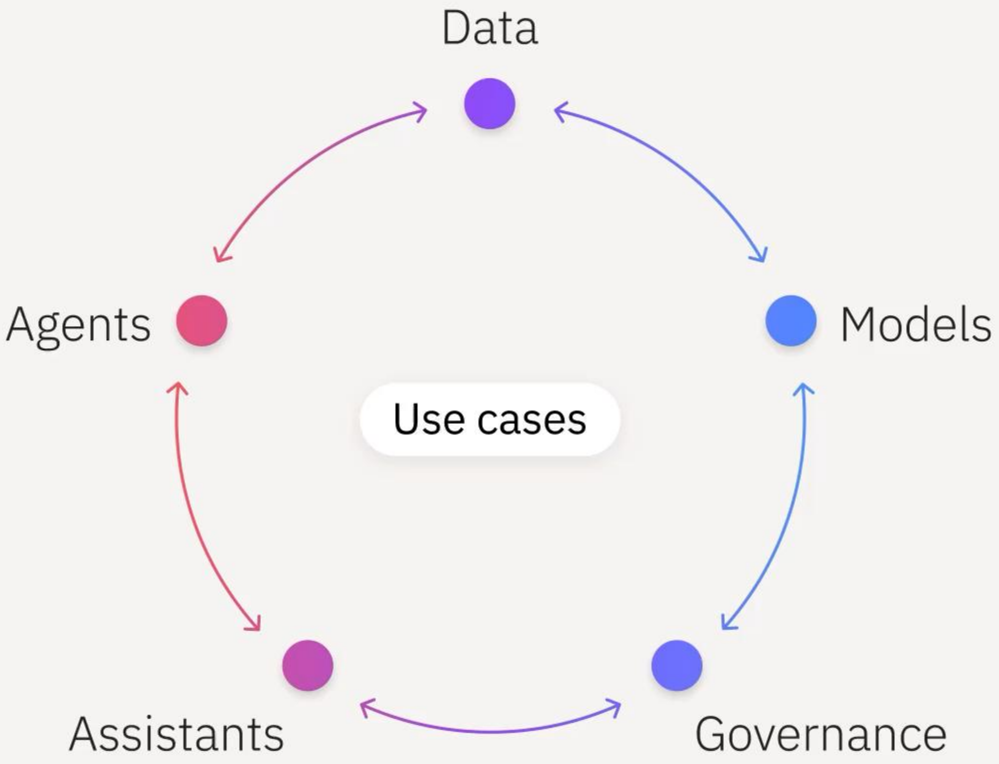
2023-2024
AI can **chat** for you

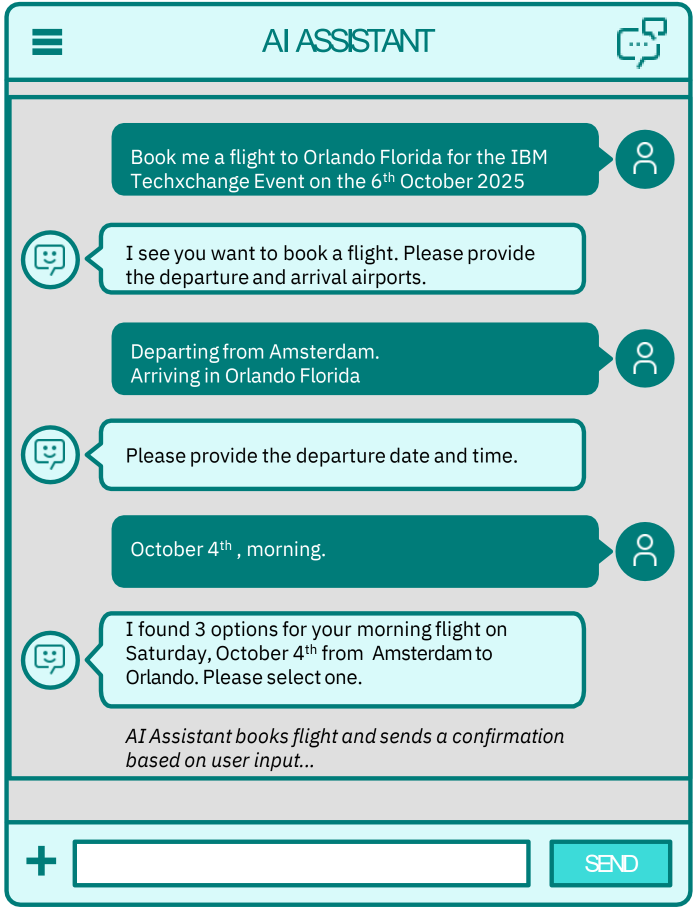
2025
AI can **do** for you!

Agentic systems can plan, take actions, reflect on the outcome



AI building blocks to the future





Book me a flight to Orlando Florida for the IBM Techxchange Event on the 6th October 2025

I see you want to book a flight. Please provide the departure and arrival airports.

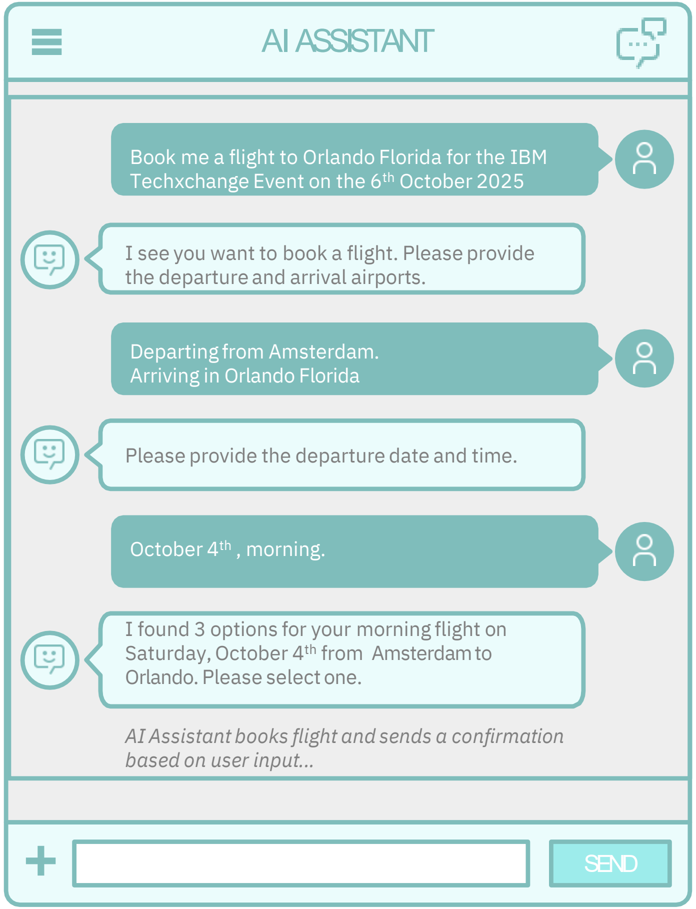
Departing from Amsterdam.
Arriving in Orlando Florida

Please provide the departure date and time.

October 4th , morning.

I found 3 options for your morning flight on Saturday, October 4th from Amsterdam to Orlando. Please select one.

AI Assistant books flight and sends a confirmation based on user input…



Book me a flight to Orlando Florida for the IBM Techxchange Event on the 6th October 2025

I see you want to book a flight. Please provide the departure and arrival airports.

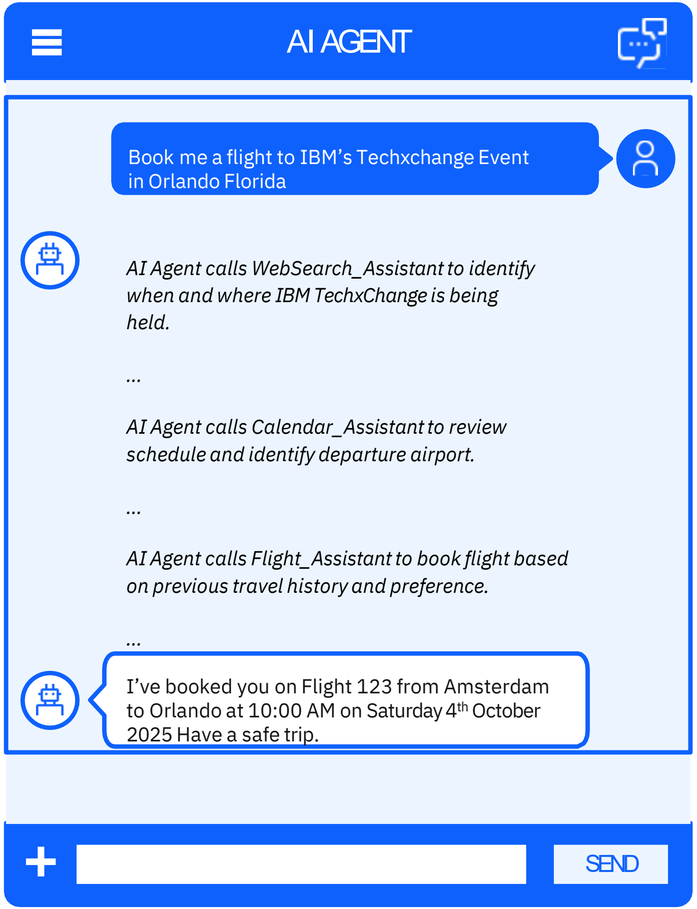
Departing from Amsterdam.
Arriving in Orlando Florida

Please provide the departure date and time.

October 4th , morning.

I found 3 options for your morning flight on Saturday, October 4th from Amsterdam to Orlando. Please select one.

AI Assistant books flight and sends a confirmation based on user input...



Book me a flight to IBM's Techxchange Event in Orlando Florida

AI Agent calls WebSearch_Assistant to identify when and where IBM TechxChange is being held.

...

AI Agent calls Calendar_Assistant to review schedule and identify departure airport.

...

AI Agent calls Flight_Assistant to book flight based on previous travel history and preference.

...

I've booked you on Flight 123 from Amsterdam to Orlando at 10:00 AM on Saturday 4th October 2025 Have a safe trip.

Understanding Risk

Rapid evolution in AI drives the need for *comprehensive* and *open* governance of that AI

Generative AI

65/35

Split in software spend on non-generative AI (65%) and generative AI (35%). [1]

50

Amplified and *new* risk with generative AI. [2]

Govern both forms of AI in a consistent manner.

Extend your governance to account for the new aspects of generative AI.

Innovation in models

~5000

New foundation models posted on Hugging Face every week.

Continuous innovation in open source and commercial offerings gives you an increasing range of options and trade-offs.

Govern the onboarding of new models.

Govern the trade-offs in use cases.

Consumption models

60/40

Split in software spend on AI platforms (60%) and AI embedded in enterprise applications (40%). [1]

Work devices

Enabled with AI.

Govern all AI, regardless of how and by whom it's created or consumed.

Legislation

100+

Articles in the EU AI Act.

10

Standardization requests in the EU AI Act.

AI Act now in force.

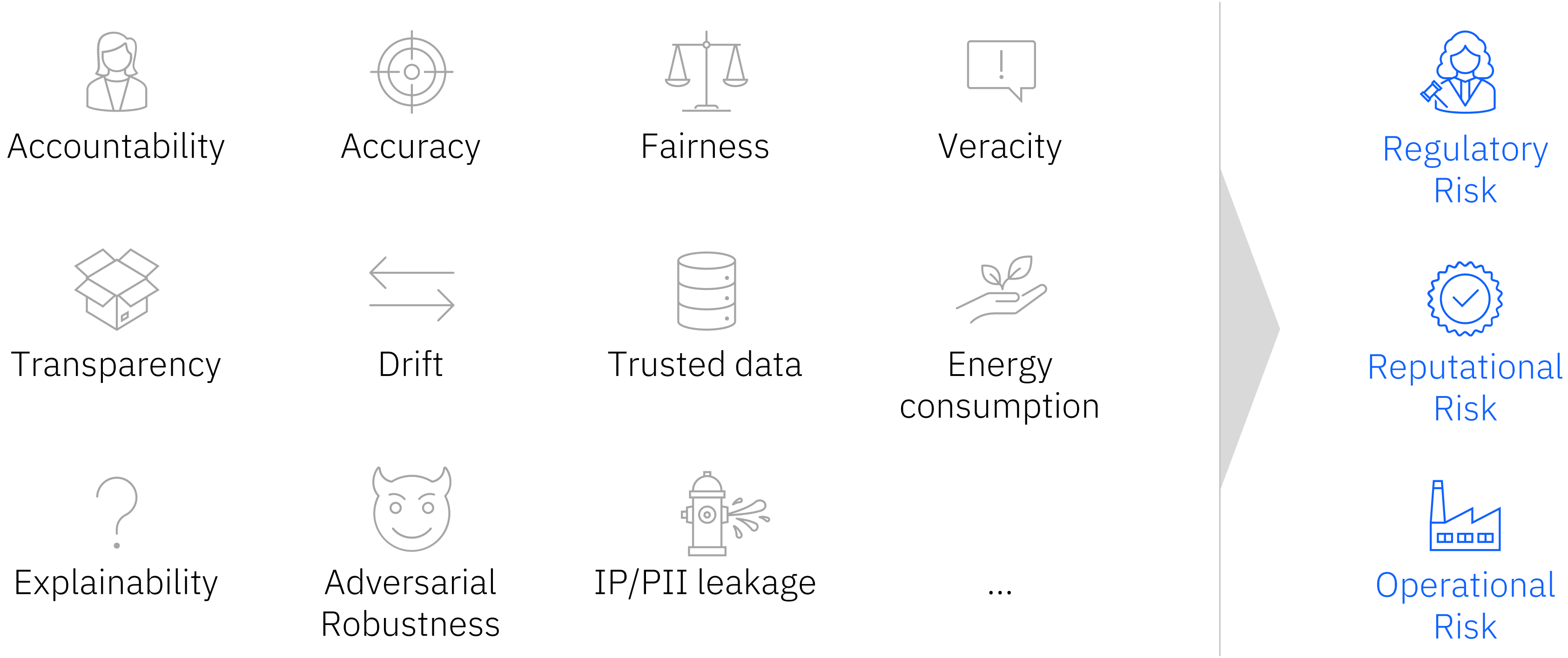
Translate your responsibilities into controls and workflows.

Risk-assess your use cases.

[1] Gartner – Forecast Analysis: AI Software. 2023-2027, Worldwide

[2] IBM AI Risk Atlas

Elements of AI risk



Accountability

Accuracy

Fairness

Veracity

Transparency

Drift

Trusted data

Energy consumption

Explainability

Adversarial Robustness

IP/PII leakage

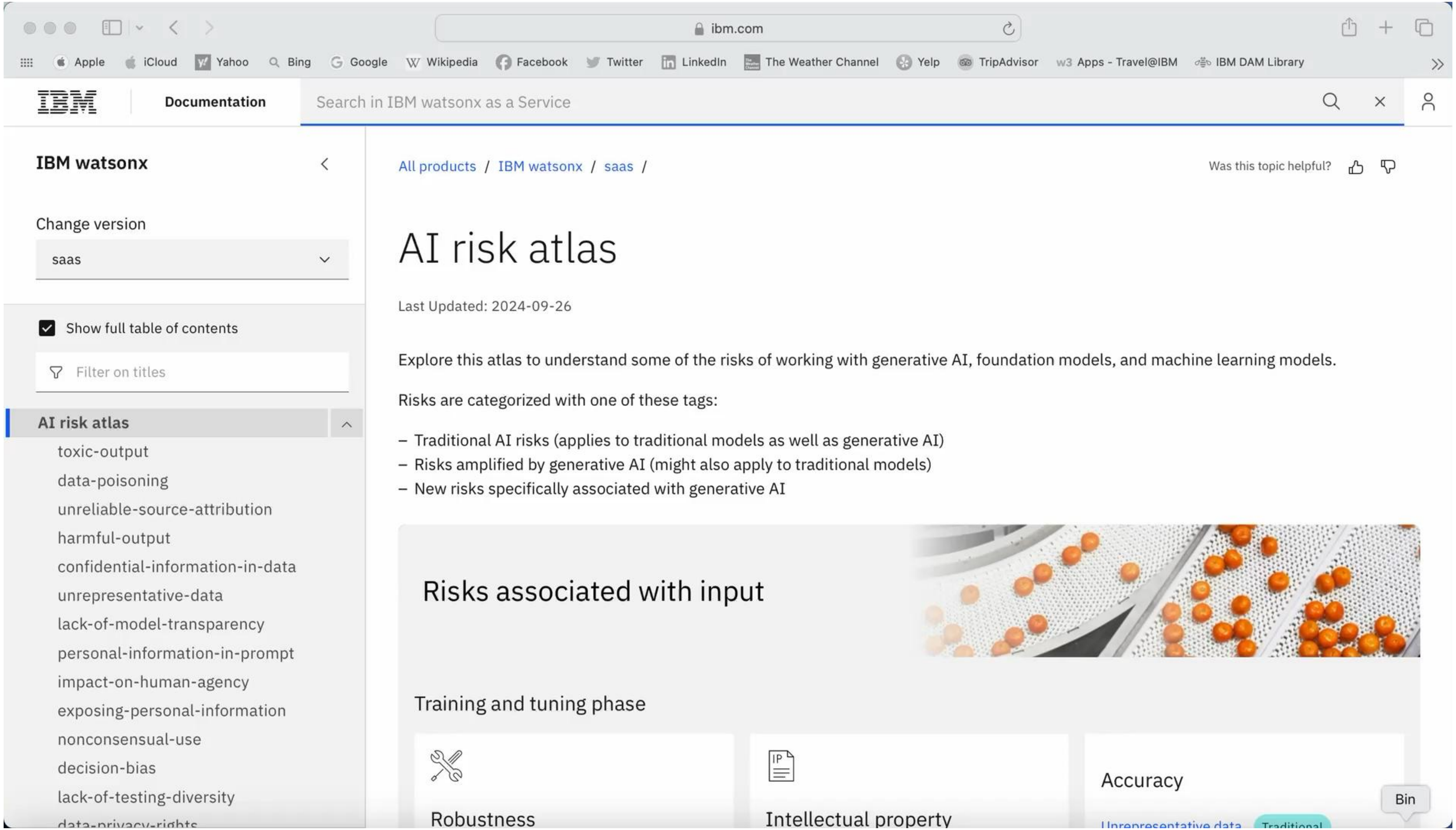
...

Regulatory Risk

Reputational Risk

Operational Risk

The IBM AI Risk Atlas



AI risk atlas
- toxic-output
- data-poisoning
- unreliable-source-attribution
- harmful-output
- confidential-information-in-data
- unrepresentative-data
- lack-of-model-transparency
- personal-information-in-prompt
- impact-on-human-agency
- exposing-personal-information
- nonconsensual-use
- decision-bias
- lack-of-testing-diversity
- data-privacy-rights

AI risk atlas

Last Updated: 2024-09-26

Explore this atlas to understand some of the risks of working with generative AI, foundation models, and machine learning models.

Risks are categorized with one of these tags:

– Traditional AI risks (applies to traditional models as well as generative AI)
– Risks amplified by generative AI (might also apply to traditional models)
– New risks specifically associated with generative AI

Risks associated with input

Training and tuning phase

Robustness

Intellectual property

Accuracy

Unrepresentative data Traditional

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

AI risk atlas - IBM Documentation

Key lifecycle governance activities

For agentic systems

Experimentation tracking

Track agentic app variants and compare results to inform which to push to production

Agentic system metrics, monitoring and alerts

Oversee elements such as hallucination, answer relevance, and system drift in production and development

Traceability

Help developers debug agentic app by tracing each step of the user interaction and agent processing

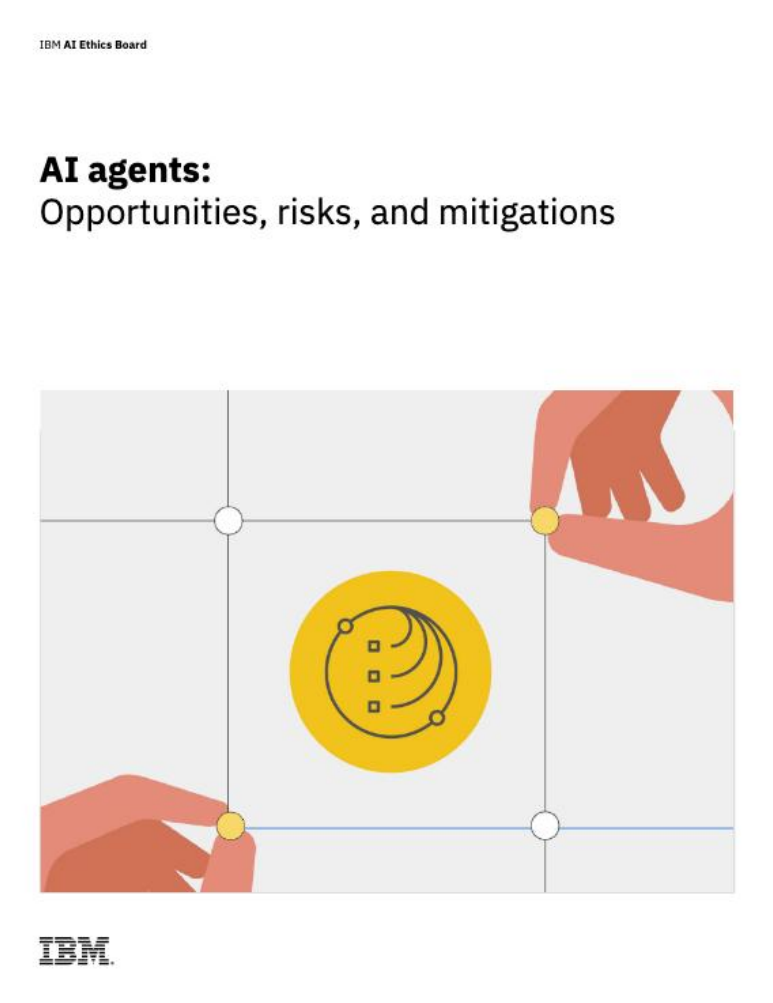
Cataloging of agentic AI applications

Single consolidated view of all in development and use

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:



IBM **AI Ethics Board**

AI agents:
Opportunities, risks, and mitigations