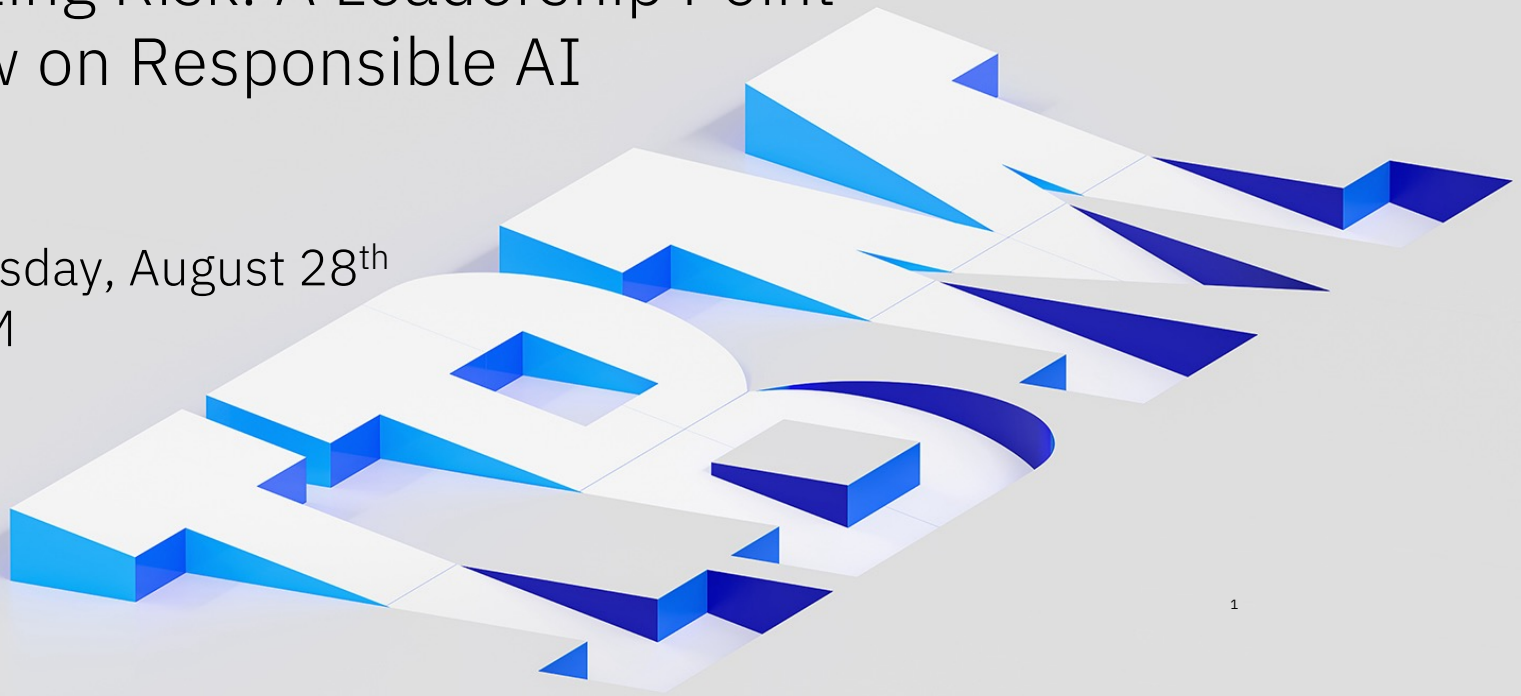
DAFITC 2024 – Montgomery, AL

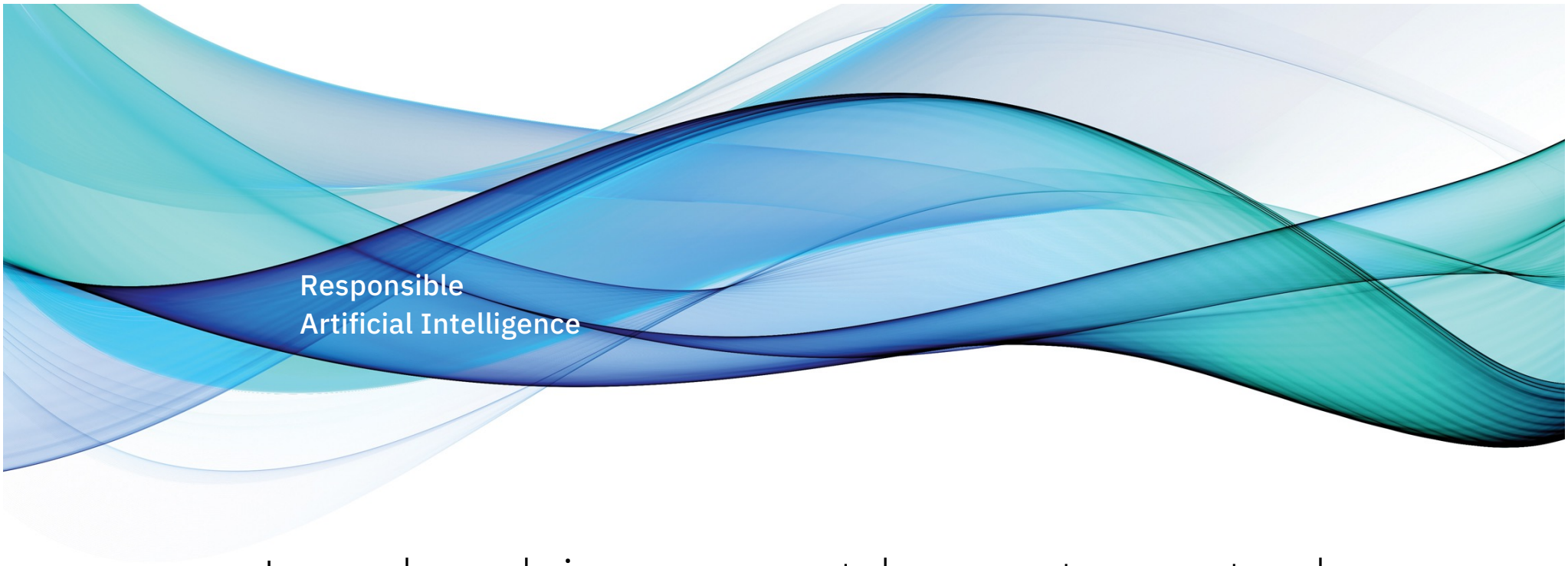
Mitigating Risk: A Leadership Point of View on Responsible AI

Wednesday, August 28th
1:50 PM



Brian Van Bibber
brian.vanbibber@us.ibm.com

IBM



Leadership cannot be automated, and ethics cannot be delegated

IBM

Brian Van Bibber, CDMP, SMB-A, MBA, BS

IBM Consulting: Federal Sector
BTS Account Lead for US Air Force and Space Force

Profile

Brian Van Bibber is a seasoned IBM consulting executive, with 11 years of data, analytics and AI consulting experience at IBM, and an additional 12+ years of prior IT and analytics experience in risk, marketing and operations analytics. He holds a Specialized Master of Business Analytics, a Master of Business Administration, and a Bachelor of Science in Computer Information Systems. He holds multiple business and technical certifications like the DAMA Certified Data Management Professional, and others which focus on data management, cloud technology and executive leadership. Brian is an Associate Partner in IBM Consulting, serving as the Business Transformation Services Focal for the United States Air Force and Space Force account.

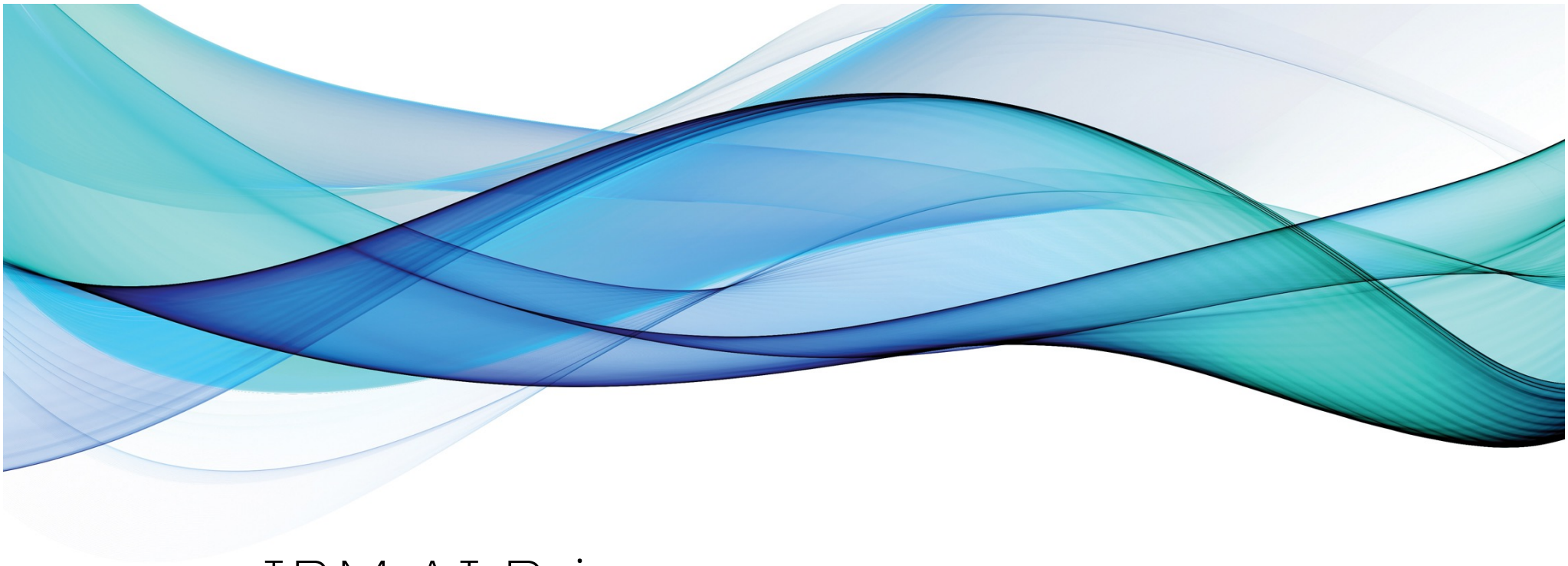
Brian's current focus is centered around the responsible use of machine learning and artificial intelligence in operations, building scalable data strategies with mesh and fabric-based platforms, data governance design and implementation, and business process modeling and automation.

Brian Van Bibber, CDMP, SMB-A, MBA
IBM Executive Consultant: Data, Analytics and AI (DAAI) Strategy & Implementation

Connect on LinkedIn

Agenda

1) AI Primer
2) Responsible AI & Risk
3) AI Ethics, Pillars and Risk Frameworks
4) Leadership Guidance on GenAI
5) Infusing Responsible AI into Operations
6) Further Reading Links
7) Questions

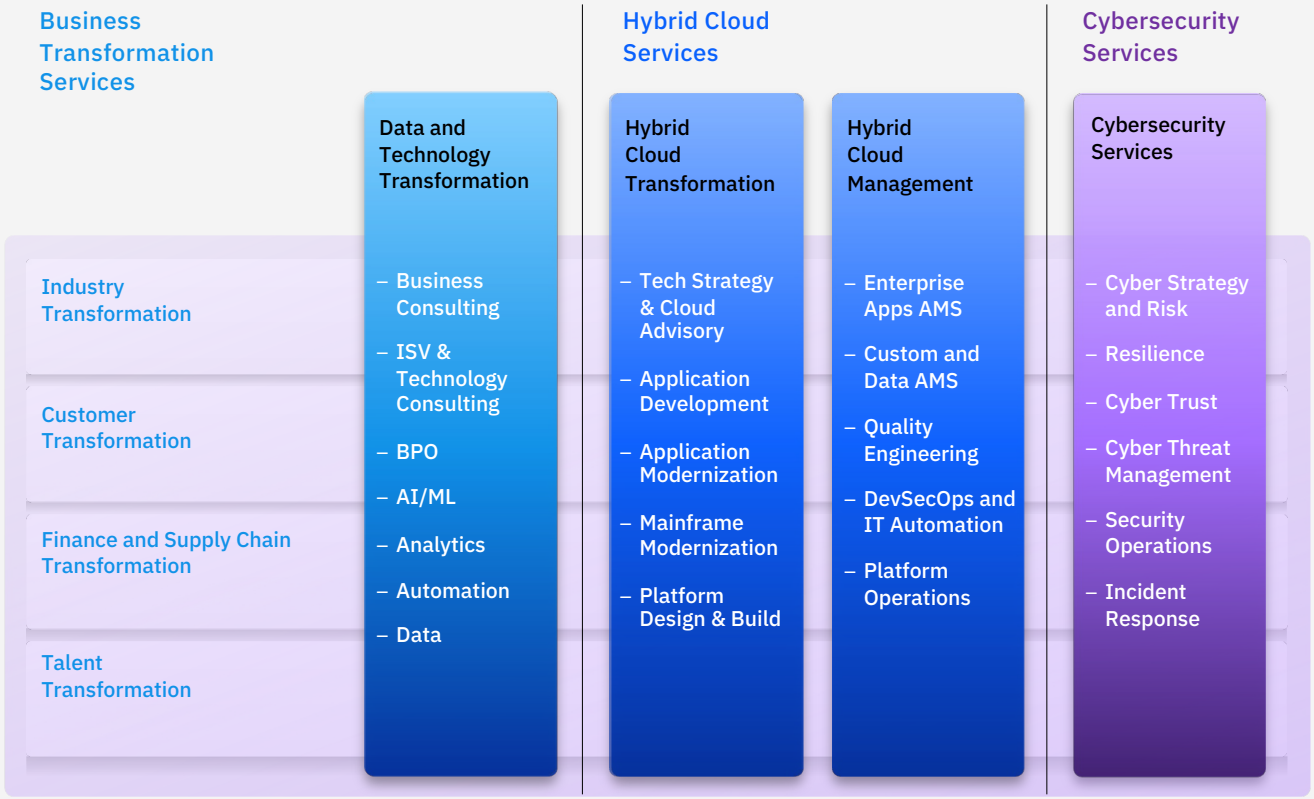


IBM AI Primer

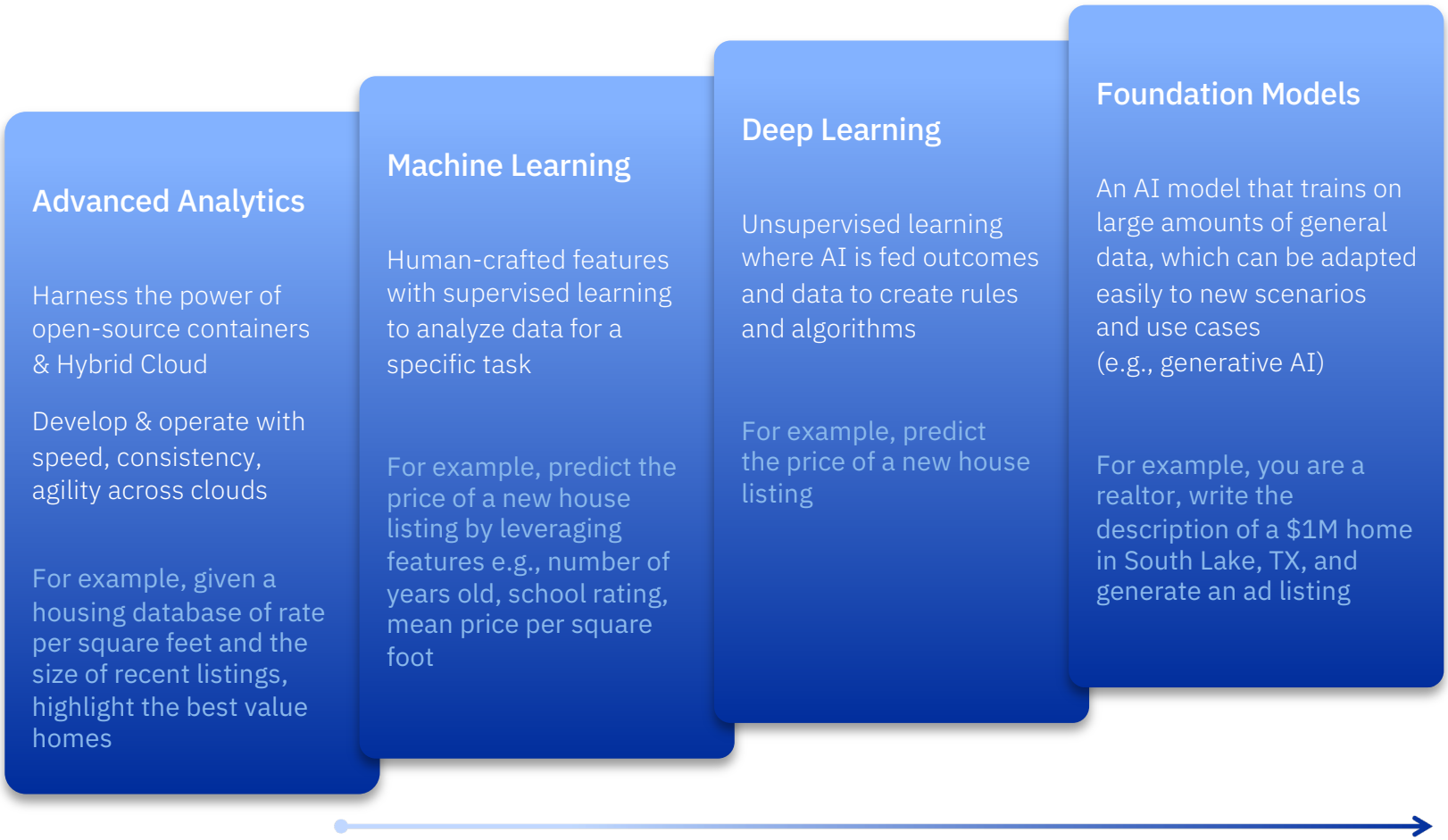
Introduction to IBM Consulting

Each service line delivers end-to-end capabilities across workflows

Our IBM Consulting team is structured around key client workflows, aligned to the full spectrum of needs of our key client buyers.



The last 2 decades have seen enormous shifts in technology



AI brings a wide spectrum of use cases..... and new risks

Traditional AI capabilities

Predictive/prescriptive
Structured data analysis, predictions, forecasting etc.

Directed conversational AI
Deterministic dialog flows for API driven conversational AI

Computer vision AI
Machine vision for object and anomaly detection

Process automation
Robotic process automation, process reengineering and optimization

Generative AI capabilities

Summarization e.g. documents such as user manuals, asset notes, financial reports, etc.

Conversational search e.g. SOPs, troubleshooting instructions, etc.

Content creation e.g. personas, user stories, synthetic data, generating images, personalized UI, marketing copy, email/social responses etc.

Code creation e.g. code co-pilot, code conversion, create technical documentation, test cases etc.

Most use cases with text/image/ video/code generation are good candidates for generative AI

Most structured data analysis, prediction and prescription are better served with traditional AI

Generative AI can augment existing traditional AI use cases to enhance natural language interactions and summarize etc.

Most common generative AI tasks today

Retrieval-augmented generation
Based on a documents or dynamic content, create a chatbot or question-answering feature.

Building a Q&A resource from a broad knowledge base, providing customer service assistance

Summarization
Transform text with domain-specific content into personalized overviews that capture key points.

Conversation summaries, insurance coverage, meeting transcripts, contract information

Content generation
Generate text content for a specific purpose.

Marketing campaigns, job descriptions, blog posts and articles, email drafting support

Named entity recognition
Identify and extract essential information from unstructured text.

Audit acceleration, SEC 10K fact extraction

Insight extraction
Analyze existing unstructured text content to surface insights in specialized domain areas.

Medical diagnosis support user research findings

Classification
Read and classify written input with as few as zero examples.

Threat and vulnerability classification, sentiment analysis, customer segmentation

Understanding Responsible AI

What are the risks that Responsible AI is designed to mitigate

Single Points of Failure

Unintended Consequences

Algorithmic Bias

Transfer of Knowledge

Data Quality / Integrity

Transparency & Explainability

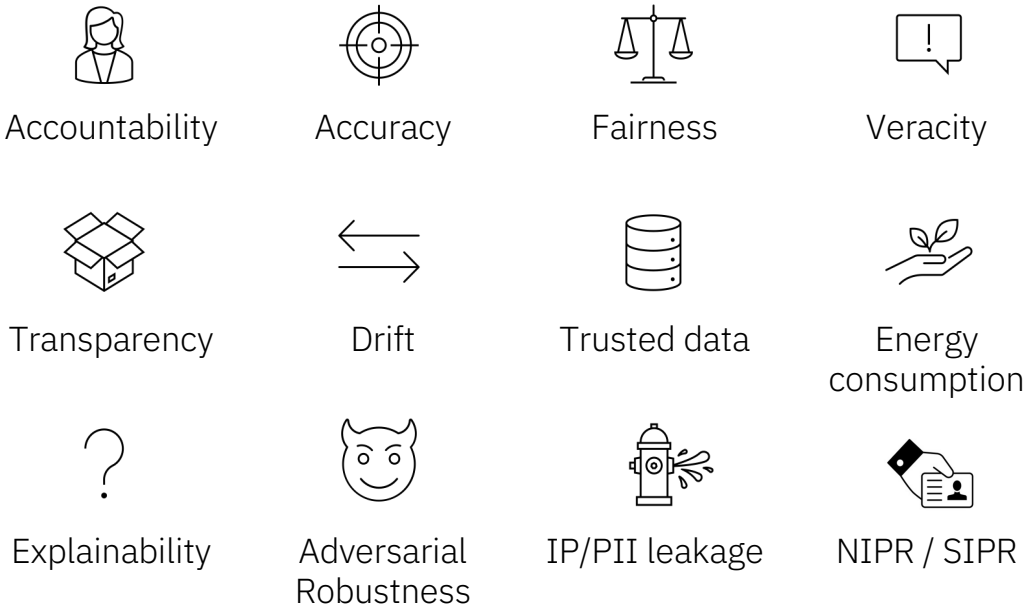
Cybersecurity

Human – AI System Collaboration

Regulatory & Policy Risks

AI System Maintenance

A foundation of ethics and principles for governance



AI Governance Components

AI Risks are grouped into three types:

Regulatory Risk

Reputational Risk

Operational Risk

AI Ethics → Responsible AI → Trustworthy AI

AI Foundation

Document Review: DoD Responsible AI Strategy

Download the Full Paper

"...directs the Department's strategic approach for operationalizing the DoD Ethical Principles and advancing RAI"

DOD AI Principles

Responsible
Judgement, care and accountability

Equitable
Minimize unintended bias

Traceable
Full transparency and documentation

Reliable
Testing and QA of capabilities

Governable
Detect and avoid unintended consequences

DOD RAI AI Strategy Tenets

Modernize Governance

Train Warfighters & Gain Trust

Mitigate Risks Across Lifecycle

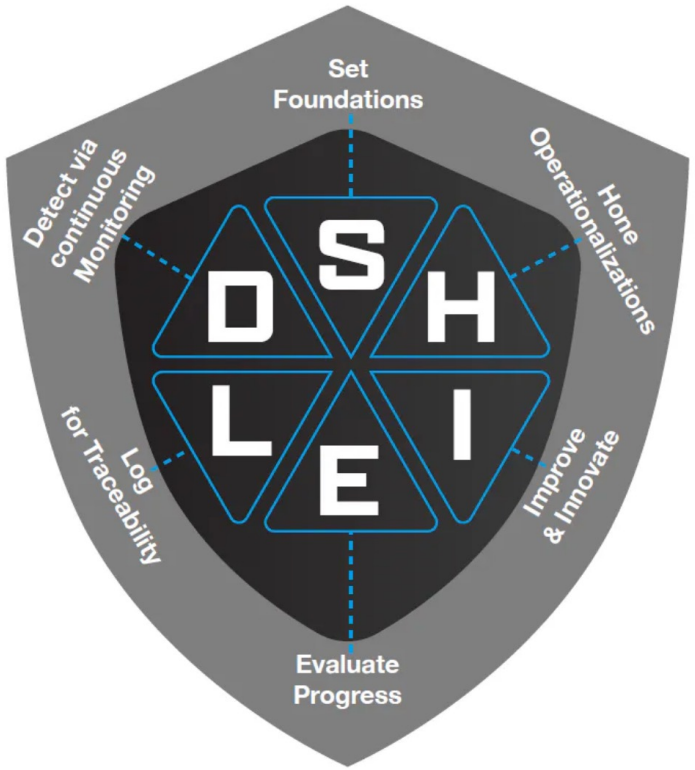
Align AI with Operations

Communicate RAI Best Practices

Deploy RAI Literacy & Gain Trust

DoD CDAO has released a Responsible AI Toolkit

View the Current RAI Tools List



RAI Toolkit Principles

Customize your solution through **Modular & Tailorable** code

Promote accountability by **Aligning to RASCI Matrix**

Approach **Holistically** to evaluate and improve AI systems

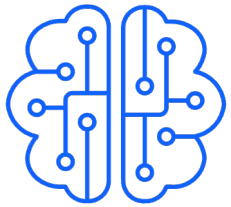
Adhere to **DoD AI Ethical Principles**

Mitigate risks & improve development with **Tools**

Building Trust and Managing Risk

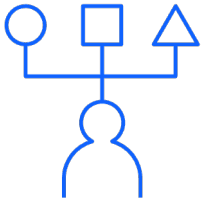
AI Ethics Principles

Download the Full Paper

01 

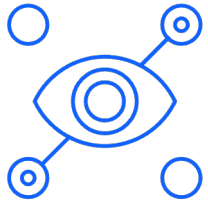
The purpose of AI is to augment—not replace—human intelligence

At IBM, we believe AI should make all of us better at our jobs, and that the benefits of the AI era should touch the many, not just the elite few.

02 

Data and insights belong to their creator

IBM clients' data is their data, and their insights are their insights. We believe that government data policies should be fair and equitable and prioritize openness.

03 

New technology, including AI systems, must be transparent and explainable

Companies must be clear about who trains their AI systems, what data was used in training and, most importantly, what went into their algorithms' recommendations.

AI Ethics Pillars

Links to IBM 360 Open-Source Toolkits

Explainability

An AI system's ability to provide a human-interpretable explanation for its predictions and insights

Is it easy to understand? Interpretable, by whom?

Fairness

Equitable treatment of individuals or groups by an AI system — depends on the context in which the AI system is used

Is it equal? Equity, meritocracy, needs-based?

Robustness

An AI system's ability to effectively handle exceptional conditions, such as abnormalities in input

Has it been tampered with by anyone?

Transparency

An AI system's ability to include and share information on how it has been designed and developed

Who is accountable? Where did the training data come from?

Privacy

An AI system's ability to prioritize and safeguard consumers' privacy and data rights

Is my data protected?

Open-Source Toolkits

AI Explainability 360

AI Fairness 360

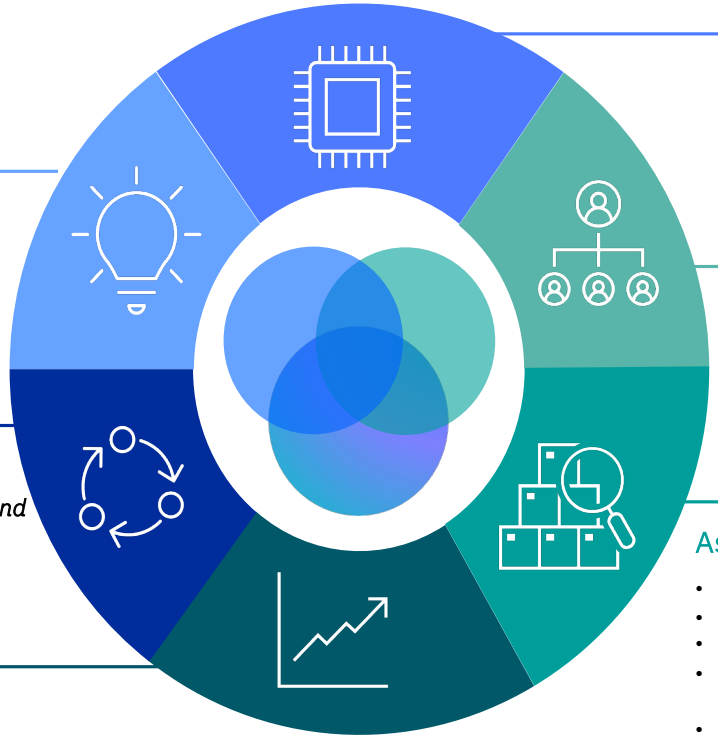
Adversarial Robustness 360

AI Factsheets 360

AI Privacy 360

Framework for Responsible Use of AI:
A Holistic Governance Approach to Mitigate Risk

Download the Full Paper



Principles, Policies & Processes

- Align to business strategy & worldview
- Establish AI Principles & Pillars aligned to values
- *Define organizational policies and processes*
- Advise on growing regulatory environment and industry best practices

Organizational structure

- *Define roles & responsibilities of personnel*
- Value Alignment
- Incentive structures
- Establish AI Ethics boards
- *Build Advocacy Network and Communities*

Assess & Audit

- The cost of a Generative AI hallucination
- *Risk assessment (initially & regularly) and mitigation*
- Consider current & upcoming Data & AI regulation
- *Review models, practices, processes, structures, skills and culture for Trustworthiness*
- Support and automate internal audits on input and output

Awareness & Training

- Benchmark learning
- *Engage and align the right stakeholders*
- Education for all levels of the business
- Co-develop a comms strategy

Operating Model & Adoption

- *Design with Ethics, Security, Privacy and Regulation in mind*
- Develop & execute playbook for Trustworthy AI
- *Guidance on implementation of AI emerging risks, and safeguards against them*

Scaled Practices

- *Scaled Data Science practices across the AI lifecycle*
- Ensure scalability and smooth operations. Maintain the business benefits of your deployed models.
- Reduce manual efforts to deploy and track model performance. Increase team efficiency
- Feedback loops and Factsheets for transparency and human agency

Identify and understand risks up front

Trustworthiness Framework for LLMs

Download the Full Paper

Input Risks

Training & Tuning

- Data Bias
- Data Poisoning
- Data Curation
- Down-stream Based Retraining
- Data Transfer
- Data Usage
- Data Acquisition
- Data Usage Rights
- Data Transparency
- Data Provenance
- PII in Data
- Reidentification of PII
- Data Privacy Rights
- Informed Consent

Inference

- PII in Prompts
- IP in Prompts
- Confidential Data in Prompts
- Evasion Attacks
- Prompt-Based Attacks

Challenges

Governance

- Model Transparency
- Accountability

Societal Impact

- Impact on Jobs
- Human Exploitation
- Environmental Impact
- Cultural Diversity
- Human Agency
- Education Impact

Legal Compliance

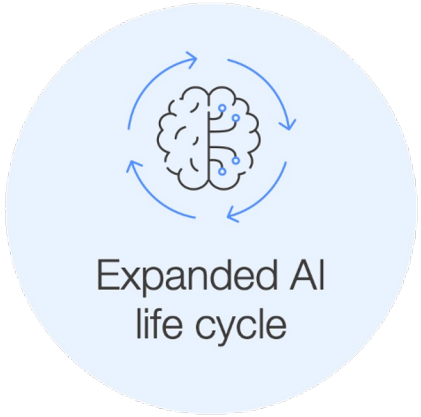
- Legal Accountability
- Content Ownership
- Content IP
- Source Attribution

Output Risks

- Output Bias
- Decision Bias
- Copyright Infringement
- Hallucinations
- Toxic Output
- Dangerous Advice
- Misinformation Spreading
- Toxicity
- Nonconsensual Use
- Dangerous Use
- Non-disclosure
- Improper Usage
- Harmful Code Generation
- Over / Under Reliance
- Exposed PII Data
- Unexplainable Output
- Unreliable Attribution of Sources

The Presidio AI Framework

Download the Full Paper



Defines the complete lifecycle of generative AI models, dividing the lifecycle into five phases:

- Data Management Phase
- Foundation Model Building Phase
- Foundation Model Release Phase
- Model Adaptation Phase
- Model Usage Phase

Implementing known and innovative guardrails ensures secure systems capable of maintaining high technical quality, uniformity, and oversight

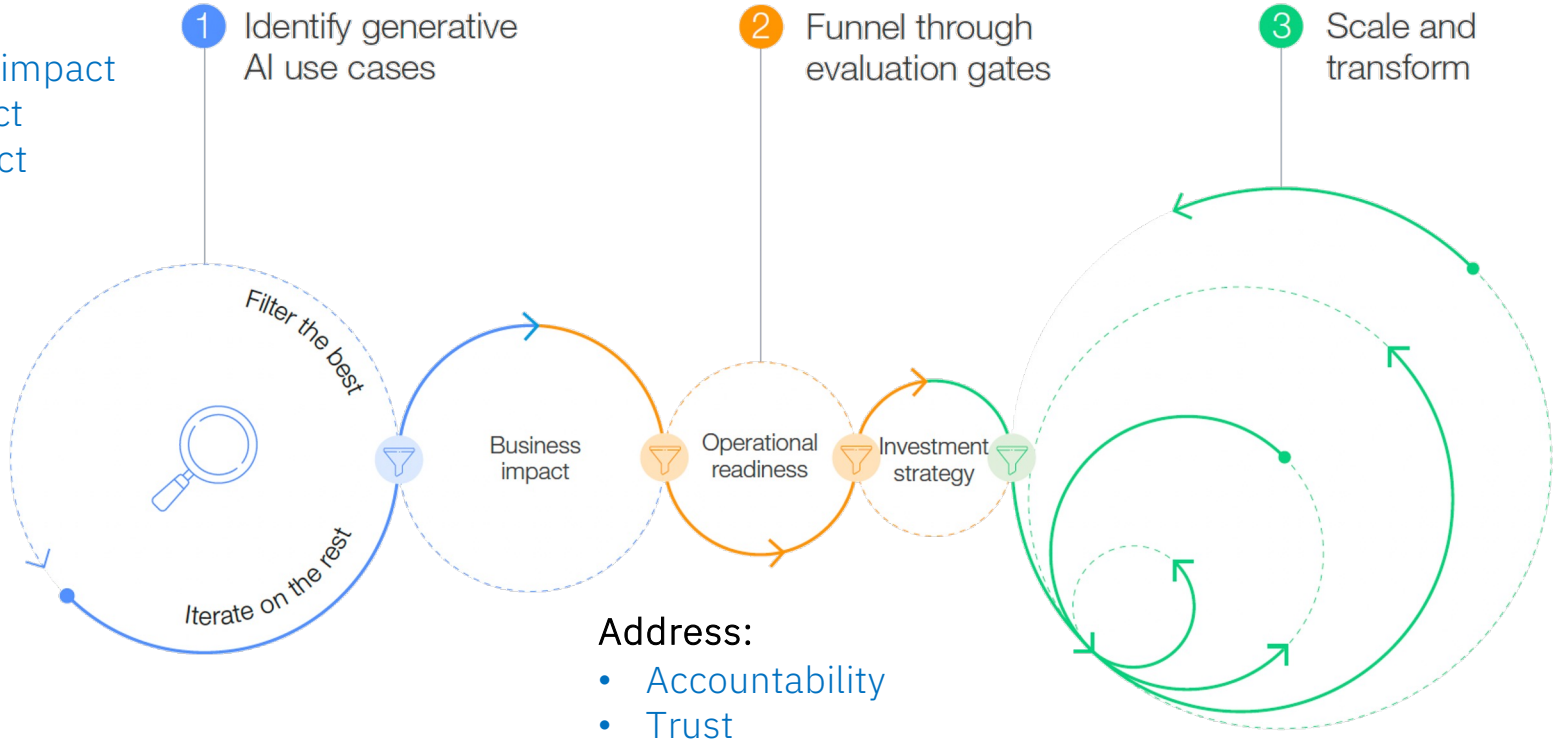
Move safety checks toward the beginning of the AI Lifecycle to minimize risks and improve overall efficiency. Early detection and resolution of problems save resources and reduce risk

Assessing Use Cases for Adoption

Download the Full Paper

Consider:

- Workforce & talent impact
- Hallucination impact
- Sustainability impact



Address:

- Accountability
- Trust
- Ability to Scale
- Change Management

Leadership and Generative AI

Get your team ready for AI

Preparing Airmen for Generative AI

Link to Document

1. Create AI-enabled People

What to know

Generative AI is all about people—how work gets done.

What to do

Make people, not technology, central to your generative AI strategy.

.

2. Set Expectations

What to know

Most leaders are overly optimistic about their organization's readiness for gen AI.

What to do

Be specific in the application of generative AI and the benefits you expect.

3. Foster Creativity

What to know

Creativity is the "must-have" generative AI skill.

What to do

Examine your operating model to unlock creativity.

Get yourself ready for AI

Build a Foundation with Responsible AI and Ethics

1. Build a Strategy

What to know

Leaders cannot pass the buck on AI ethics.

What to do

Give ethics teams a seat at the table

.

2. Align on Trust

What to know

All decisions are being scrutinized, do not jeopardize on trust.

What to do

Earn trust by aligning with user expectations

3. Ambiguity & Compliance

What to know

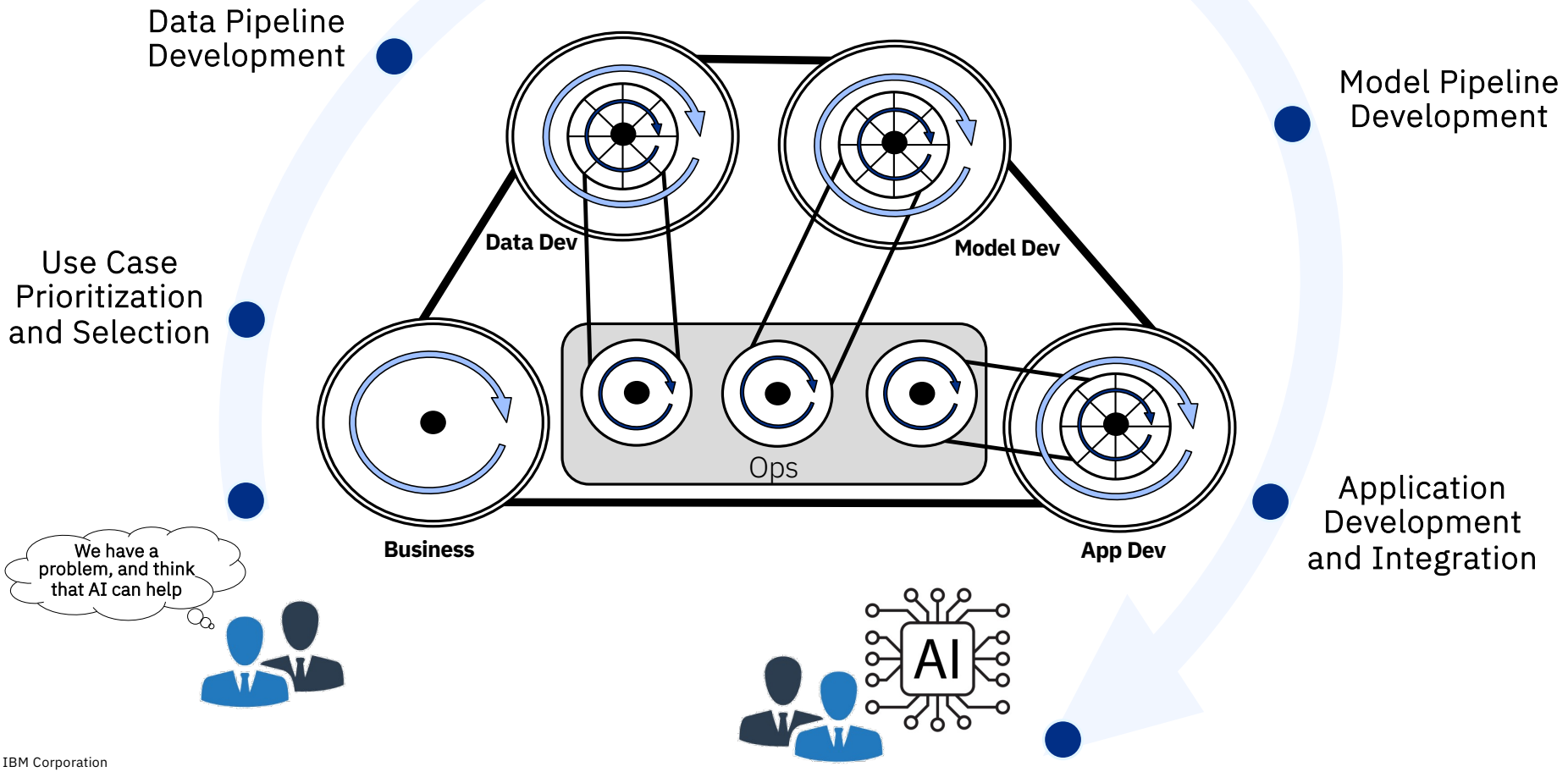
Some organizations freeze in the headlights of regulatory ambiguity

What to do

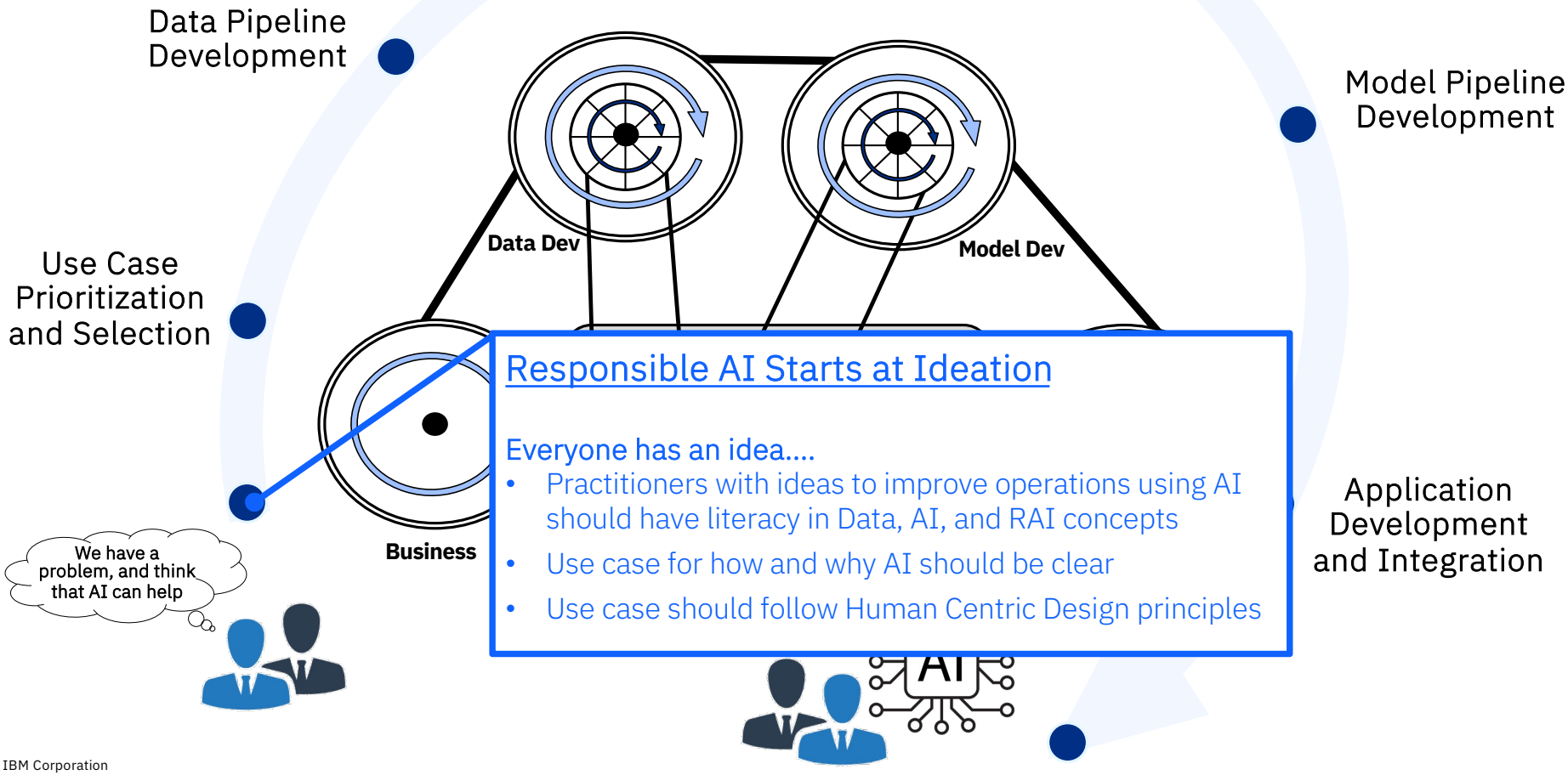
Bake in ethics and regulatory preparedness for all AI and data investments

Infusing Responsible AI into Operations

Responsible AI Infused into the Operations Framework



Responsible AI Infused into the Operations Framework

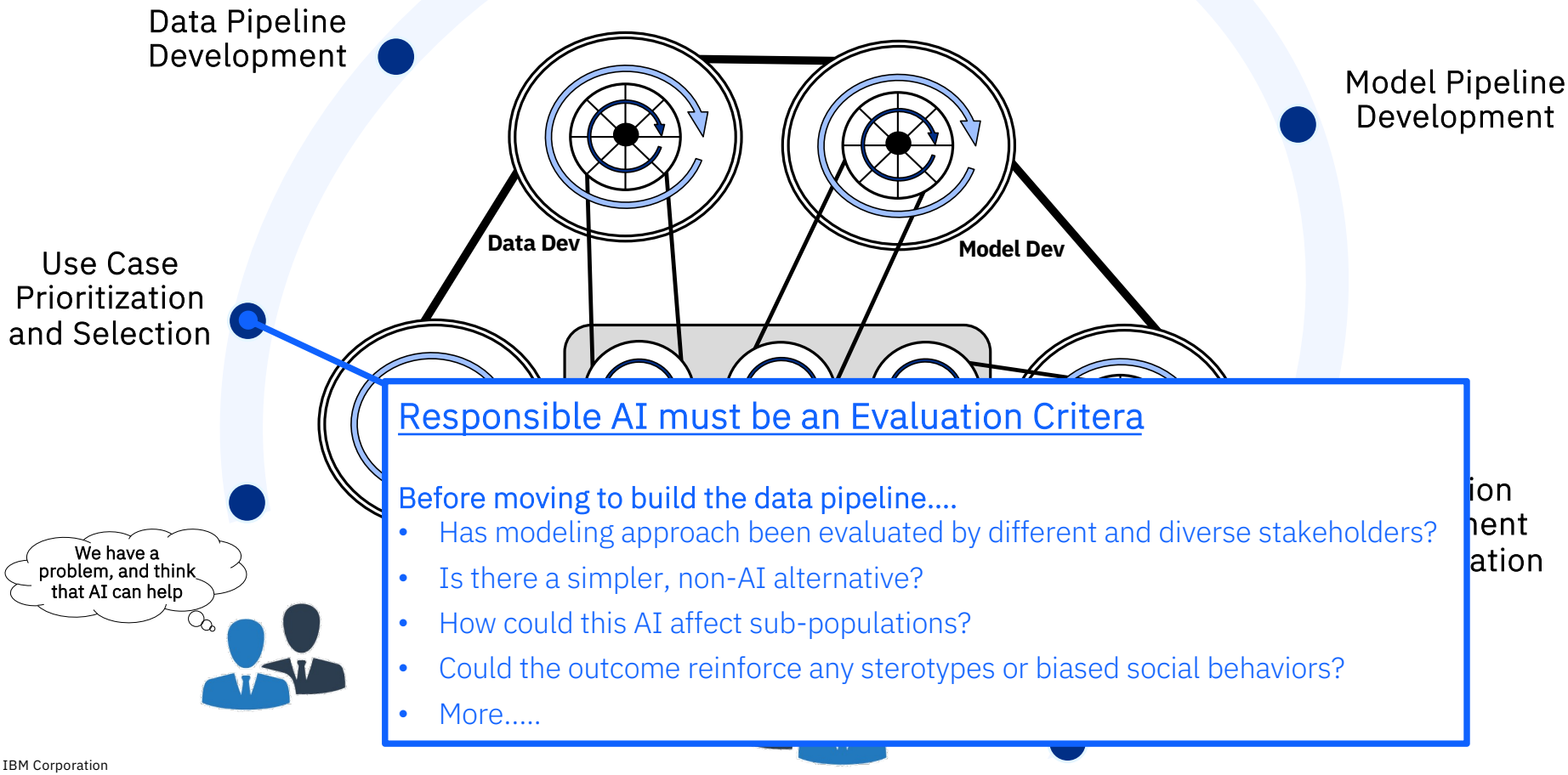


Responsible AI Starts at Ideation

Everyone has an idea....

- Practitioners with ideas to improve operations using AI should have literacy in Data, AI, and RAI concepts
- Use case for how and why AI should be clear
- Use case should follow Human Centric Design principles

Responsible AI Infused into the Operations Framework

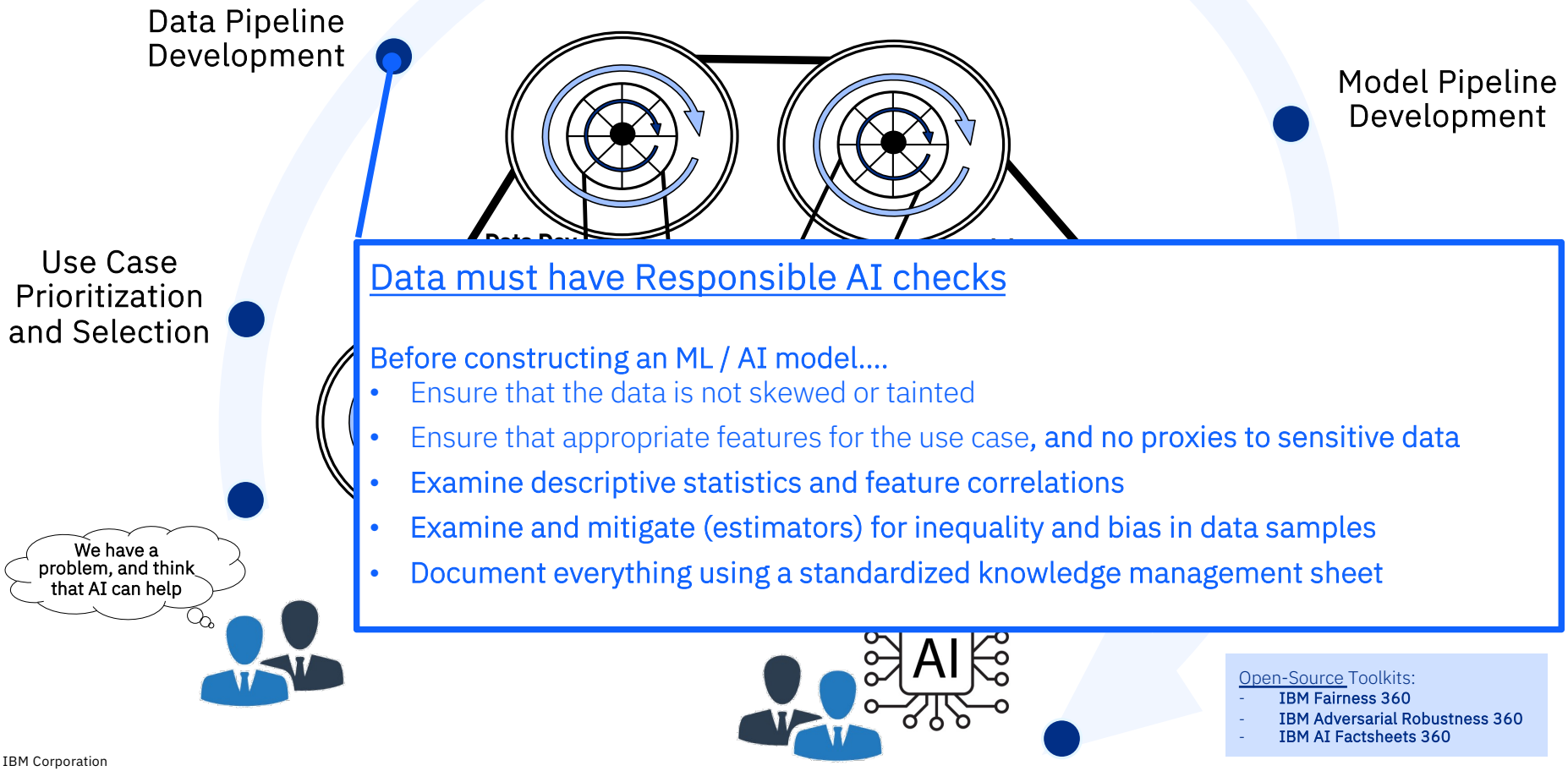


Responsible AI must be an Evaluation Critera

Before moving to build the data pipeline....

- Has modeling approach been evaluated by different and diverse stakeholders?
- Is there a simpler, non-AI alternative?
- How could this AI affect sub-populations?
- Could the outcome reinforce any sterotypes or biased social behaviors?
- More.....

Responsible AI Infused into the Operations Framework



Data must have Responsible AI checks

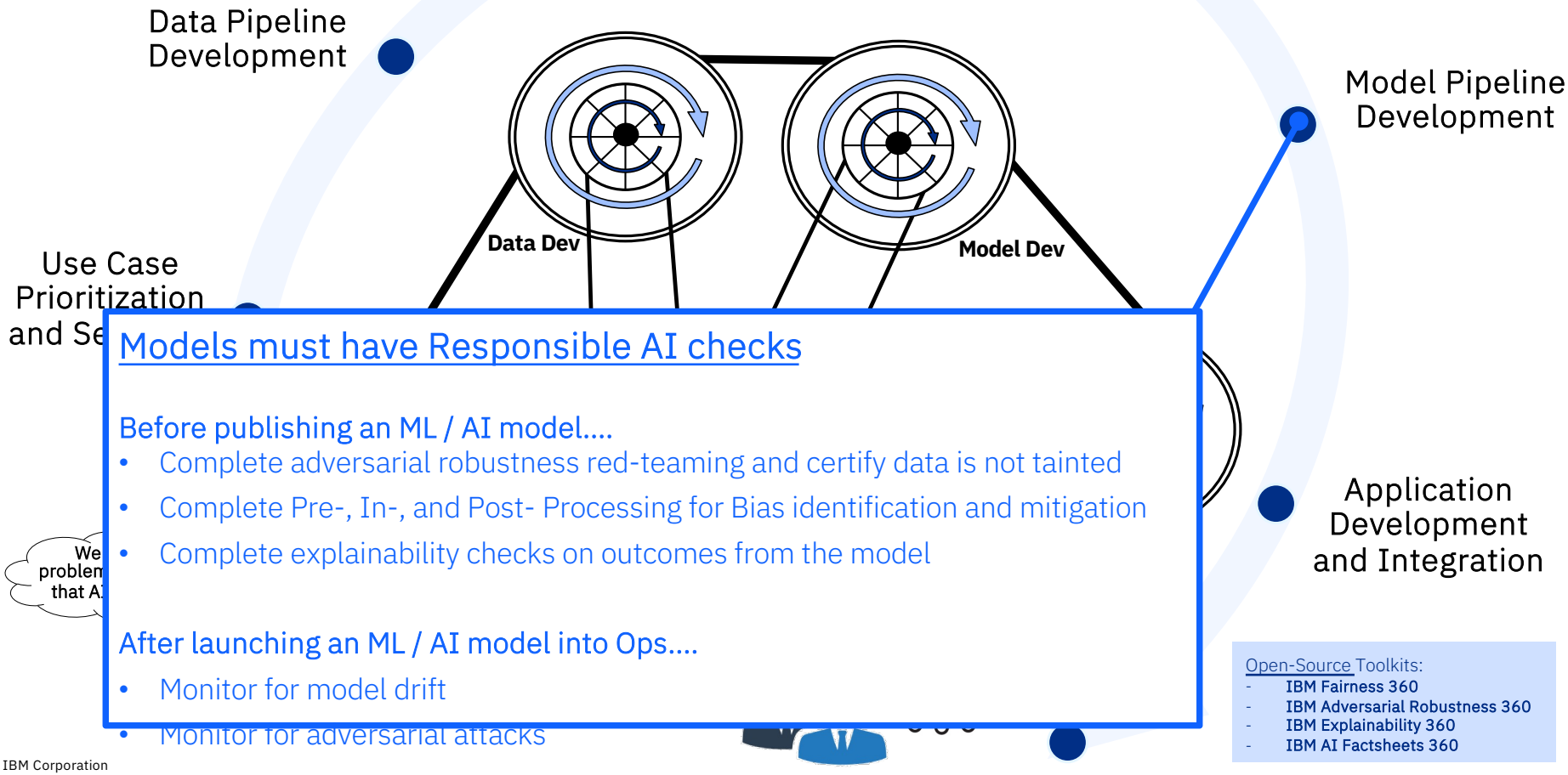
Before constructing an ML / AI model....

- Ensure that the data is not skewed or tainted
- Ensure that appropriate features for the use case, and no proxies to sensitive data
- Examine descriptive statistics and feature correlations
- Examine and mitigate (estimators) for inequality and bias in data samples
- Document everything using a standardized knowledge management sheet

Open-Source Toolkits:

- IBM Fairness 360
- IBM Adversarial Robustness 360
- IBM AI Factsheets 360

Responsible AI Infused into the Operations Framework



Models must have Responsible AI checks

Before publishing an ML / AI model....

- Complete adversarial robustness red-teaming and certify data is not tainted
- Complete Pre-, In-, and Post- Processing for Bias identification and mitigation
- Complete explainability checks on outcomes from the model

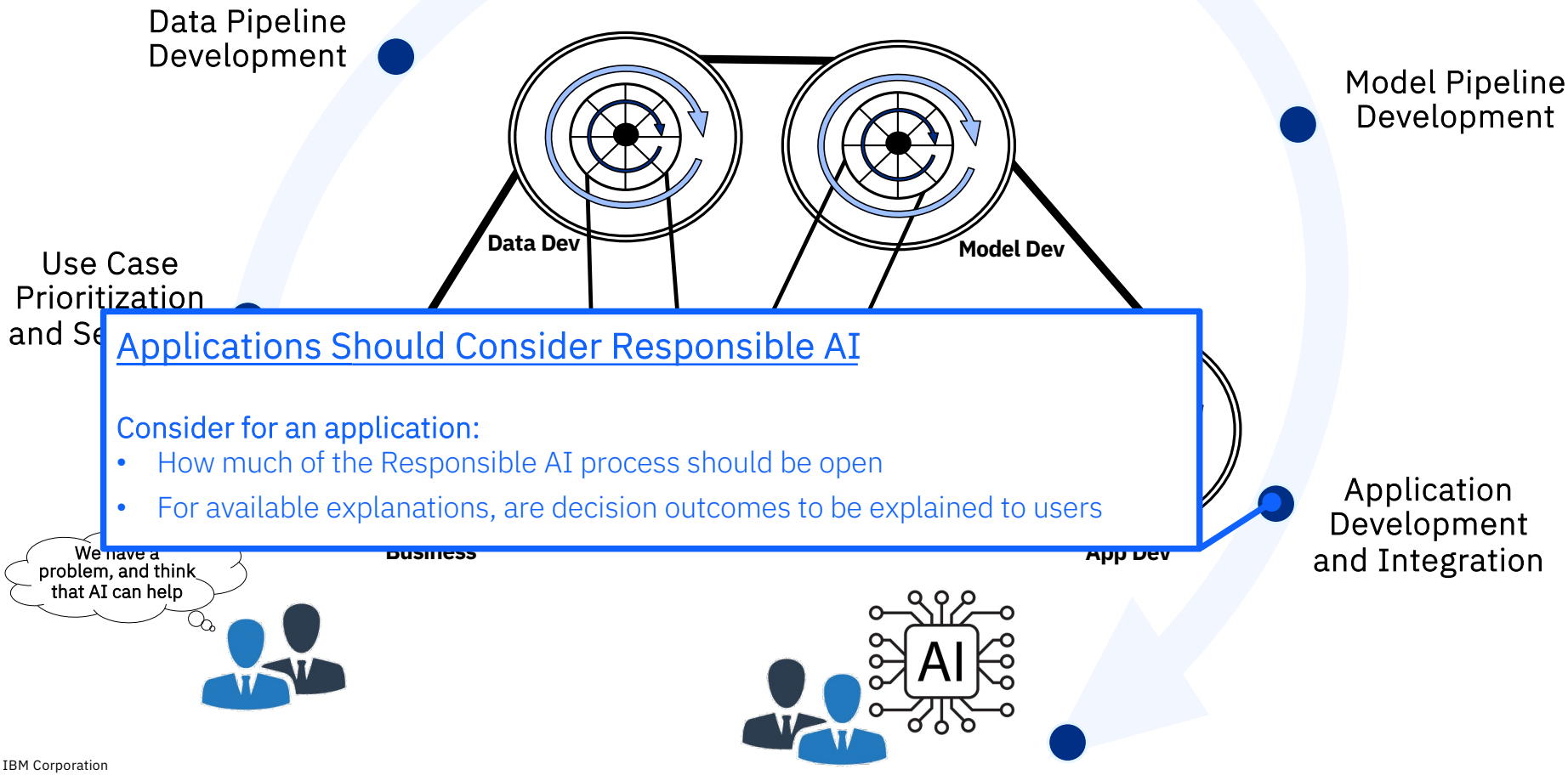
After launching an ML / AI model into Ops....

- Monitor for model drift
- Monitor for adversarial attacks

Open-Source Toolkits:

- IBM Fairness 360
- IBM Adversarial Robustness 360
- IBM Explainability 360
- IBM AI Factsheets 360

Responsible AI Infused into the Operations Framework

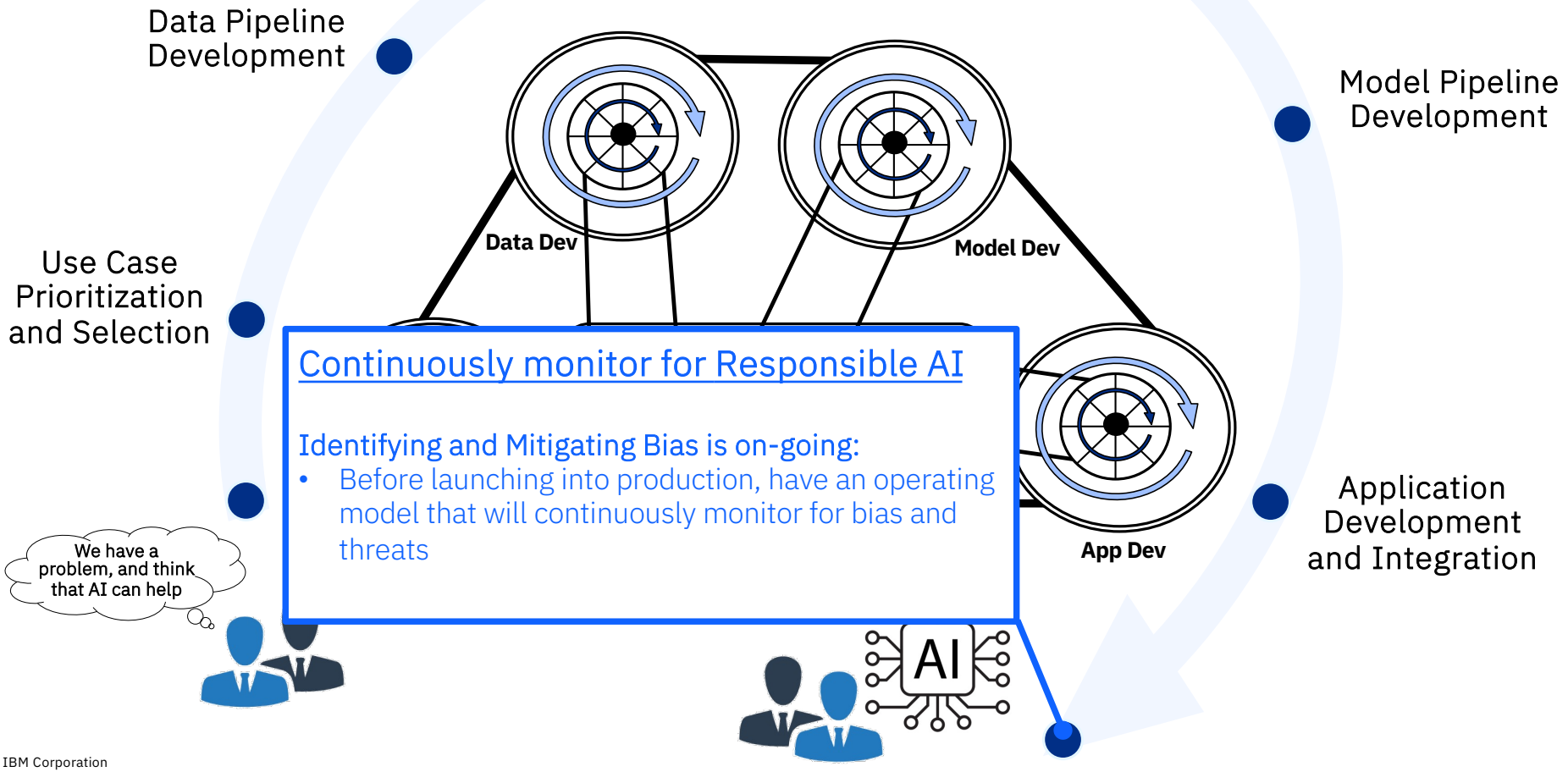


Applications Should Consider Responsible AI

Consider for an application:

- How much of the Responsible AI process should be open
- For available explanations, are decision outcomes to be explained to users

Responsible AI Infused into the Operations Framework



Thank You!

Schedule a Follow Up Meeting
with the IBM Team

Extra Links & Documentation

Policies, positions, and points of view

Actively working with regulators and industry leaders to shape AI governance and regulations

Precision regulation

[How governments and companies should advance trusted AI]

Chief Executive Officer Arvind Krishna discusses three core tenets for smart AI regulation.

[US Senate Testimony]

Chief Privacy and Trust Officer Christina Montgomery's testimony before the US Senate Judiciary Committee on a precision regulation approach to AI.

[Precision Regulation for AI]

IBM believes that companies should utilize a risk-based AI governance policy framework and targeted policies to develop and operate trustworthy AI.

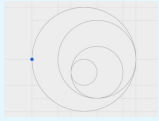
[Precision Regulation for Data-Driven Business Models]

White paper outlining seven recommendations about data-driven business model risks for policymakers.

[Precision Regulation and Facial Recognition]

IBM's position on a precision regulation approach to facial recognition that strikes the right balance between unlocking benefits and mitigating legitimate concerns.

Generative AI and foundation models



[Foundation models: Opportunities, risks and mitigations]

IBM's in-depth perspective on the opportunities posed by foundation models as well as potential risks and mitigation techniques.

[A Policymaker's Guide to Foundation Models]

IBM's recommendations to policymakers for taking proactive, productive steps toward addressing unique concerns related to foundation models.

Responsible AI and technology

[Mitigating Bias in Artificial Intelligence]

[Standards for Protecting At-Risk Groups in AI Bias Auditing]

[Promoting the Responsible Advancement of Neurotechnology]

[Data Responsibility @ IBM]

IBM's Trustworthy AI Governance: Tools

IBM has developed a suite of products & open source toolkits to enable effective AI governance

OPEN SOURCE TOOLKITS

AI Explainability 360
Comprehensive open-source toolkit for explaining ML models & data.

AI Fairness 360
Comprehensive open-source toolkit for detecting & mitigating bias in ML models.

Adversarial Robustness 360
Comprehensive open-source toolkit for defending AI from attacks.

AI FactSheets 360
A research effort to foster trust in AI by increasing transparency and enabling governance.

AI Privacy 360
Toolbox to support the assessment of privacy risks of AI-based solutions, and to help them adhere to any relevant privacy requirements.

Uncertainty Quantification 360
Comprehensive open-source toolkit for computing and communicating meaningful limitations of ML predictions.

IBM PRODUCT OFFERINGS

IBM watsonx
New enterprise-ready AI and data platform comprised of three powerful components: watsonx.ai, watsonx.data, and watsonx.governance.

watsonx.governance
Enables monitoring large language models throughout their lifecycle, proactively managing AI risk and tracing AI decision-making, supports regulatory compliance and helps clients prepare for future regulatory standards and requirements

IBM Cloud Pak for Data
A data and AI platform with a data fabric that makes data available for AI and analytics, on any cloud, and supports the creation of AI FactSheets.

IBM Watson Studio
Empowers data scientists, developers, and analysts to build, run, and manage AI models, and optimize decisions anywhere on IBM Cloud Pak for Data.

IBM Watson OpenScale
Tracks and measures outcomes from AI throughout its lifecycle, adapts and governs AI in changing business situations, and monitors AI models for bias, fairness, and trust.