Guardrails for Safe and Scalable GenAI Programs

Jean Liu, Executive Director, Liu Pursuits
Doug Graham, Director of Enterprise Data Governance, Mercy
Ren-Yi Lo, Head of Autonomous Systems & Data Governance, Siemens Healthineers, AI Tech Center
Brian Rasquinha, Associate Director, Solution Architecture, Privacy Analytics (Moderator)

November 14, 2025

PRIVACY ANALYTICS
an IQVIA company

Part I: Background

Who are the Executive Advisory Board?

- 21 senior-level executives from **public and private**-sector organizations
- Geographies: US, Canada, UK, EU
- Industries: healthcare, life sciences, financial services

- Operates as a forum for exchanging insights on **safe and responsible use** of sensitive data to drive innovation

- Identified a need for additional guidance for AI:
  - Evaluating the risk of use-cases
  - Establishing appropriate guardrails for use.

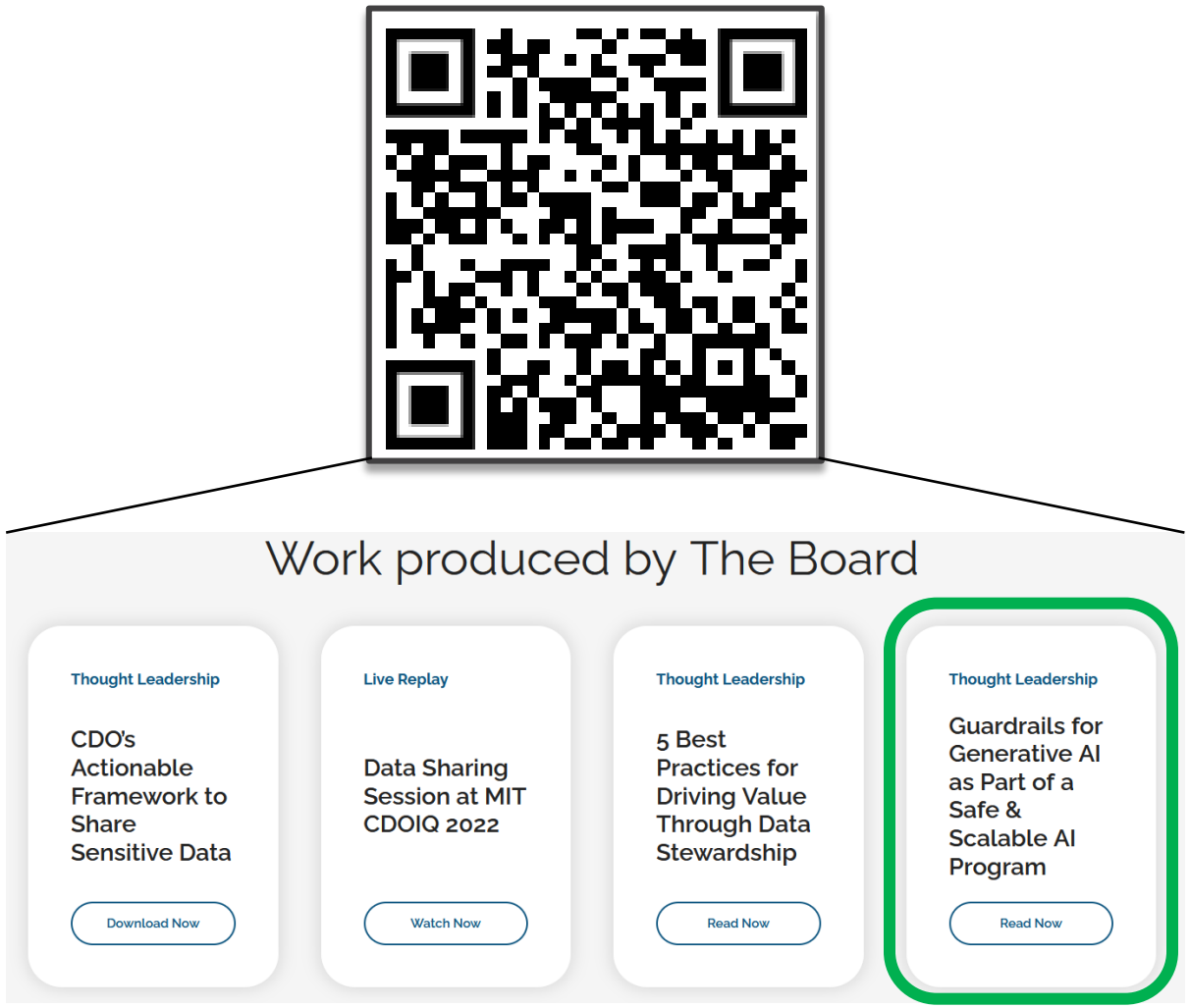
PRIVACY ANALYTICS
an IQVIA company

How was the guidance developed?

- Guidance was developed over the course of several quarterly meetings, leveraging the cross-regulation, cross-industry experience of board members

- Overall goal is to drive public and partner trust, while supporting organizations in building efficiency and responding to opportunities for innovation.

Guidance available here:



PRIVACY ANALYTICS
an IQVIA company

How was the guidance developed?

- Intended for risks manageable at the **institutional level;** other risks may be more appropriately managed by governmental or intergovernmental organizations and are outside the scope of the guidance.
- Intended to contribute to risk **mitigation**
  - Does not claim to result in minimal achievable risk
  - Does not imply any persistent residual risk is unacceptable
- Not prescriptive or exhaustive and does not constitute legal advice

- Given the advancements in AI, assessment of risk and choice of appropriate guardrails should be continually reviewed rather than a one-and-done effort.

PRIVACY ANALYTICS
an IQVIA company

Part II: Guidance

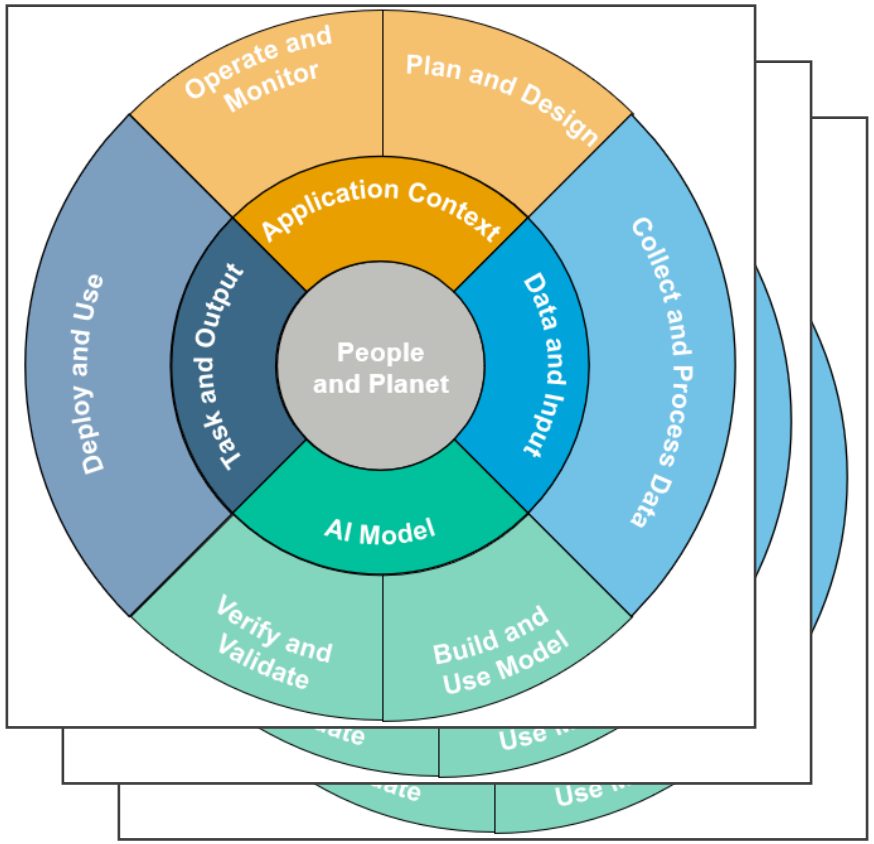
PRIVACY ANALYTICS
an IQVIA company

Guidance Overview

1. Risk Level for Use Cases

	Lower Risk				Higher Risk
Context of Users	Under contract	Internal to organization		Internal to partnered organization	External to organization
Context of Impact	Internal			External	Social, economic, professional; e.g., digital therapy
Transparency to users, regulators	Human in the loop, output checking	AI-generated content is tagged		Use of AI is explicitly stated	
Purposes	Reference (e.g., summaries, categorization)	Research		Decision support	Decision-making
Organizational Risk	Operational impacts		Business impacts		Legal, regulatory impacts

1. Guardrails mapped to NIST AI Risk Management Framework (3 Dimensions)



PRIVACY ANALYTICS
an IQVIA company

1. Evaluating Risk Level for Use Cases across Multiple Dimensions

- Dimensions affecting the **organizational risk** associated with GenAI use case

- Features/factors are presented from “Lower Risk” to “Higher Risk”

- **All dimensions** are intended to be taken into consideration
  - Overall risk level is intended to reflect a consolidated posture

Component

PRIVACY ANALYTICS
an IQVIA company

1. Evaluating Risk Level for Use Cases across Multiple Dimensions

	Lower Risk				Higher Risk
Context of Users	Under contract	Internal to organization		Internal to partnered organization	External to organization
Context of Impact	Internal			External	Social, economic, professional; e.g., digital therapy
Transparency to users, regulators	Human in the loop, output checking		AI-generated content is tagged	Use of AI is explicitly stated	
Purposes	Reference (e.g., summaries, categorization)		Research	Decision support	Decision-making
Organizational Risk	Operational impacts		Business impacts		Legal, regulatory impacts

PRIVACY ANALYTICS
an IQVIA company

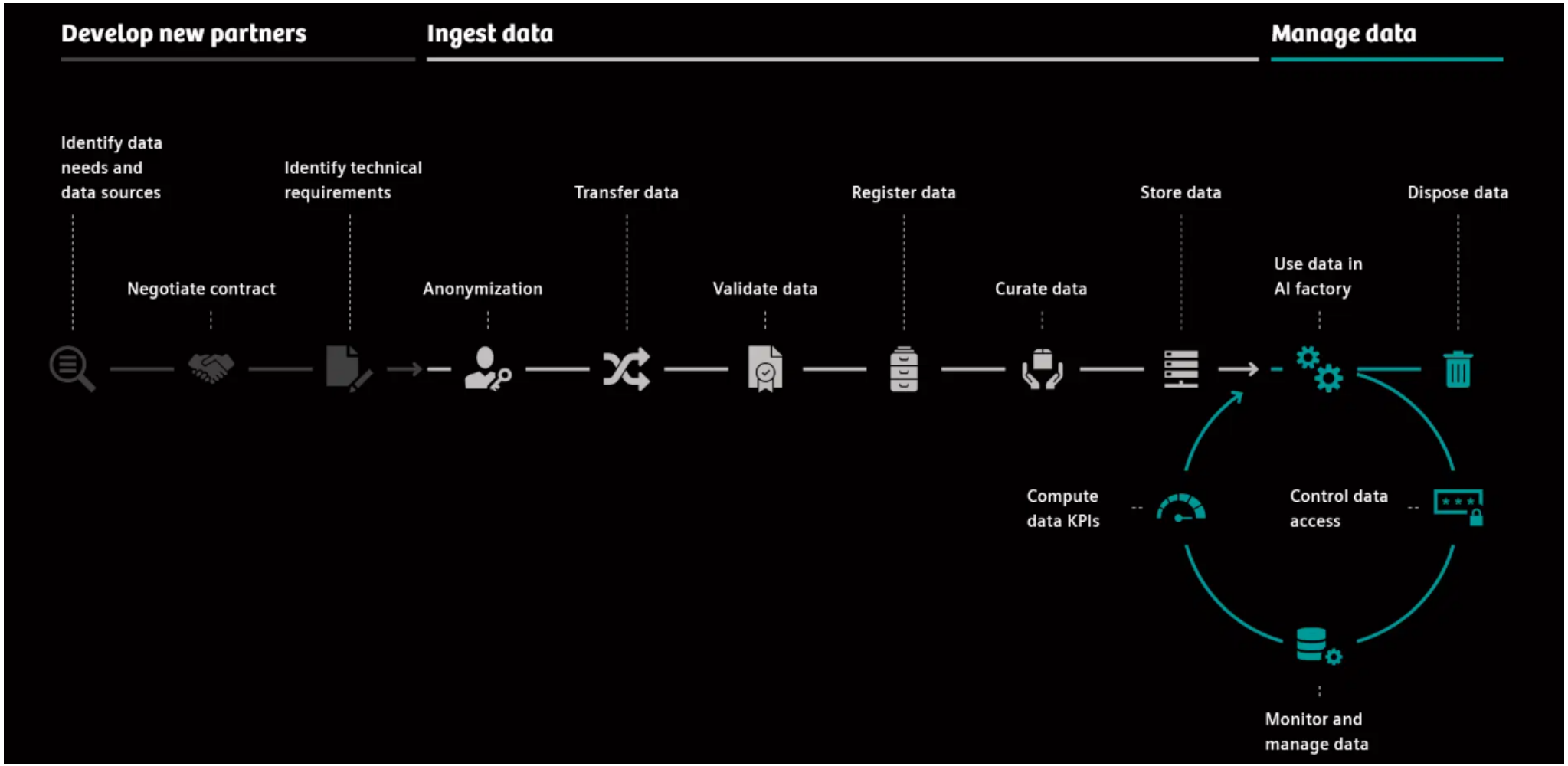
Example: Evaluating Risk Level

- **Mercy's GEN-AI SBAR Handoff Process**
- Mercy Health has implemented a **Generative AI-powered SBAR (Situation, Background, Assessment, Recommendation)** workflow for ED-to-IP transitions:
  - **Trigger Point:** When an ED admission order is placed and an inpatient bed is assigned.
  - **Automation:** GEN-AI compiles ED provider notes, clinician docs, and discrete patient data.
  - **Structured Output:** Creates a concise SBAR summary.
  - **Delivery:** Sent via Epic secure chat to the inpatient nursing team; notifications in Rover & Epic.
  - **Continuous Updates:** The AI-generated note updates until patient transfer is complete.
  - **Benefits:**
    - › Reduces cognitive burden on ED nurses.
    - › Improves accuracy and minimizes omissions.
    - › Enhances throughput by reducing phone calls and manual steps.
  - **Provider Role:** Timely ED documentation is critical for SBAR completeness.

Dimensions
User Context: – Internal (L)
Impact: – Internal (L)
Transparency: – HITL (L)
Purpose: – Decision Support (M)
Org Risk: – Business Impact (M)

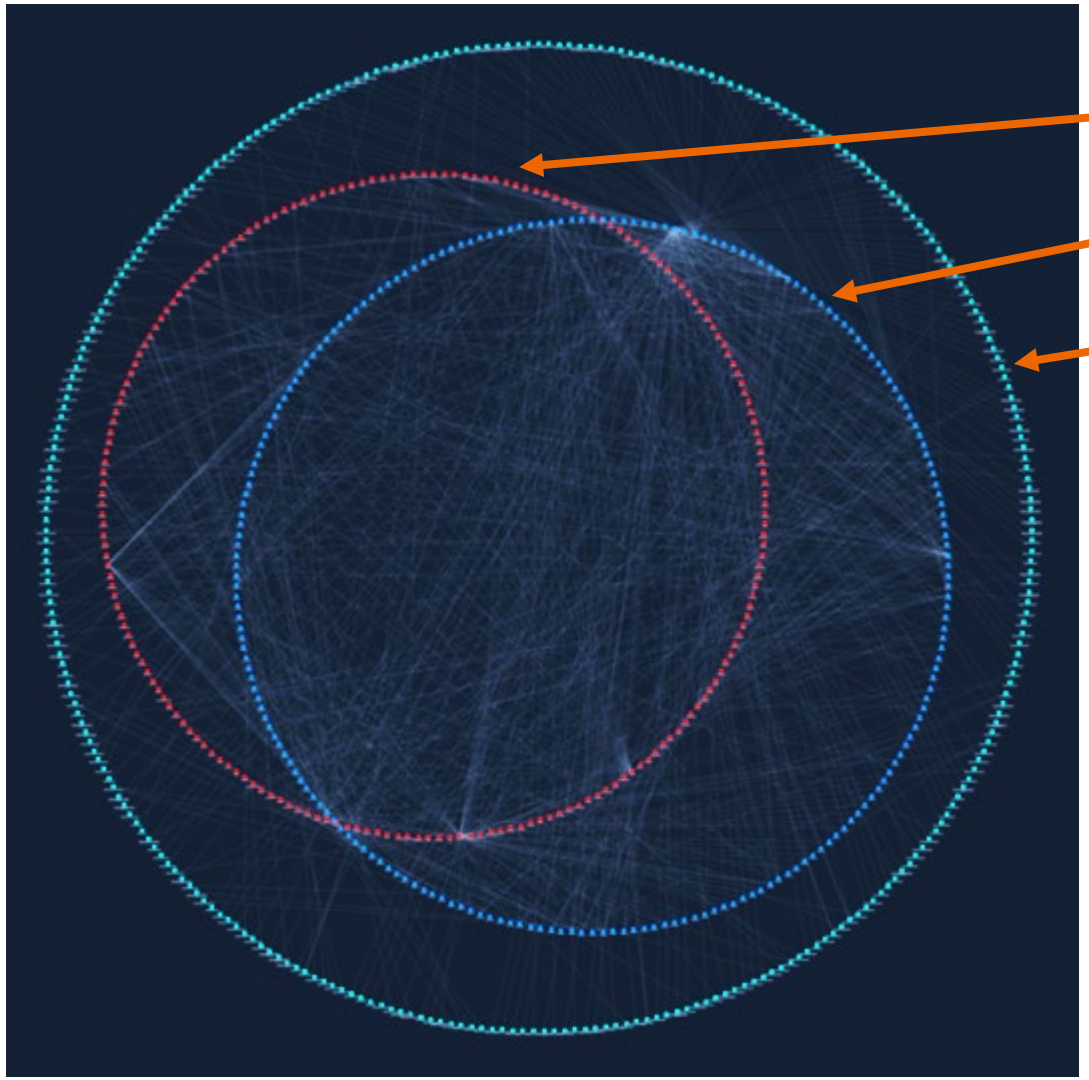
PRIVACY ANALYTICS
an IQVIA company

Comprehensive Data Lifecycle Management for Trustworthy AI Development



Data Collections, AI Projects, Researchers

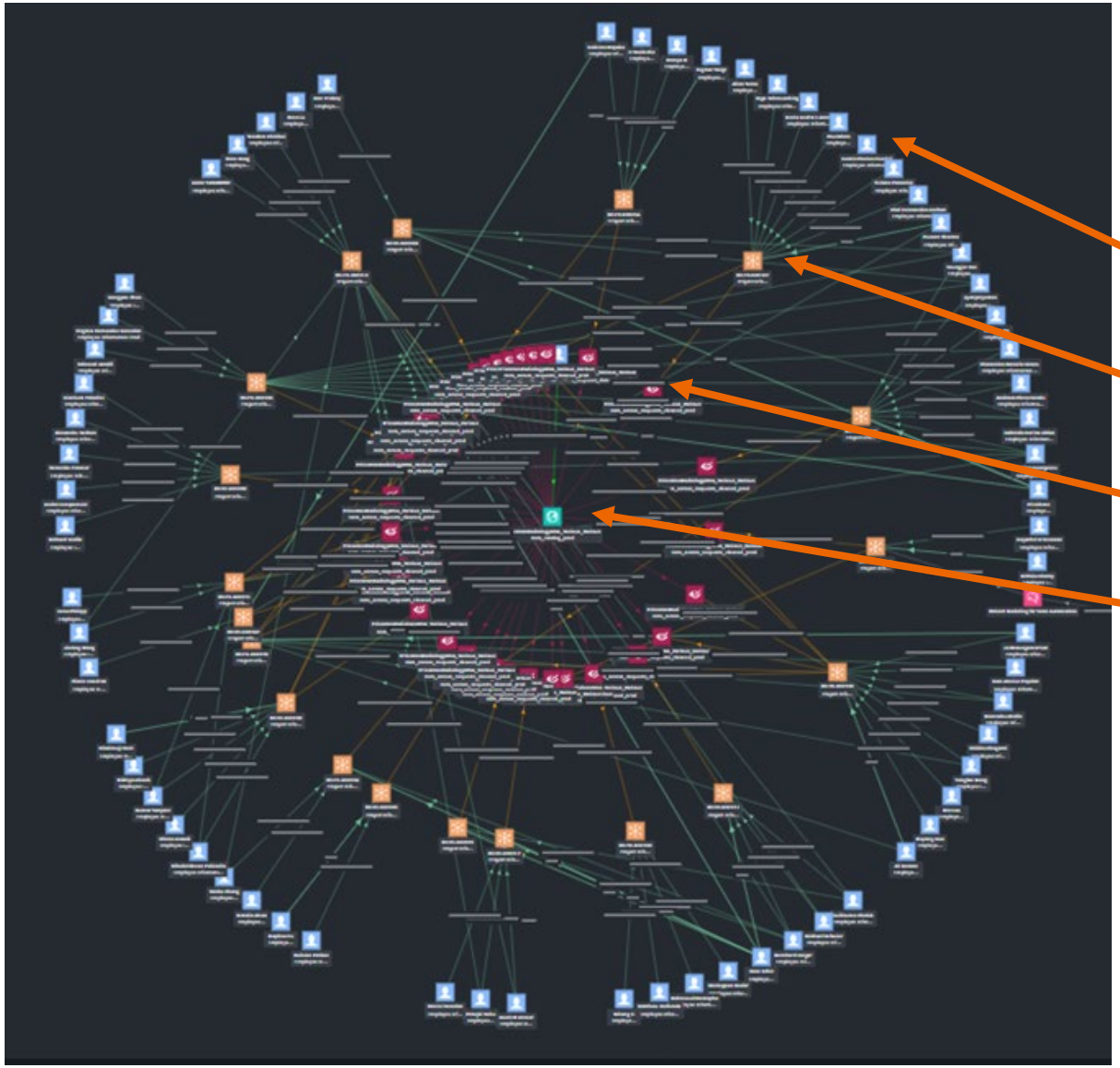
SIEMENS Healthineers



PRIVACY ANALYTICS
an IQVIA company

How Data Access and AI Projects are Managed

SIEMENS Healthineers



Researcher

AI Project

Data Access

Data Collection

PRIVACY ANALYTICS
an IQVIA company

Automating data governance extraction for data management

SIEMENS Healthineers

Problem Statement:

Can our Data Management Control Center automate extraction of data governance from data contracts by utilizing GenAI solution?

Business Case:

Efficiency Gains: from 4h to 10min

Risk Evaluation:

- Results are consumed internally to manage AI and Data governance
- Human-in-the-loop towards Human-on-the-loop
- Additional guardrails: sufficient high-quality historical data, domain experts, audits

PRIVACY ANALYTICS
an IQVIA company

2. Guardrails Mapped to NIST AI Risk Management Framework

- Guardrails mapped to the NIST AI Risk Management Framework (AI RMF)

- The guardrails are further broken down into three categories of approaches:
  - Safety & Adaptability
  - Governance & Transparency
  - Trust & Ethical Responsibility

- Higher-risk use cases -> usually stronger, more numerous, and/or broader guardrails

Component

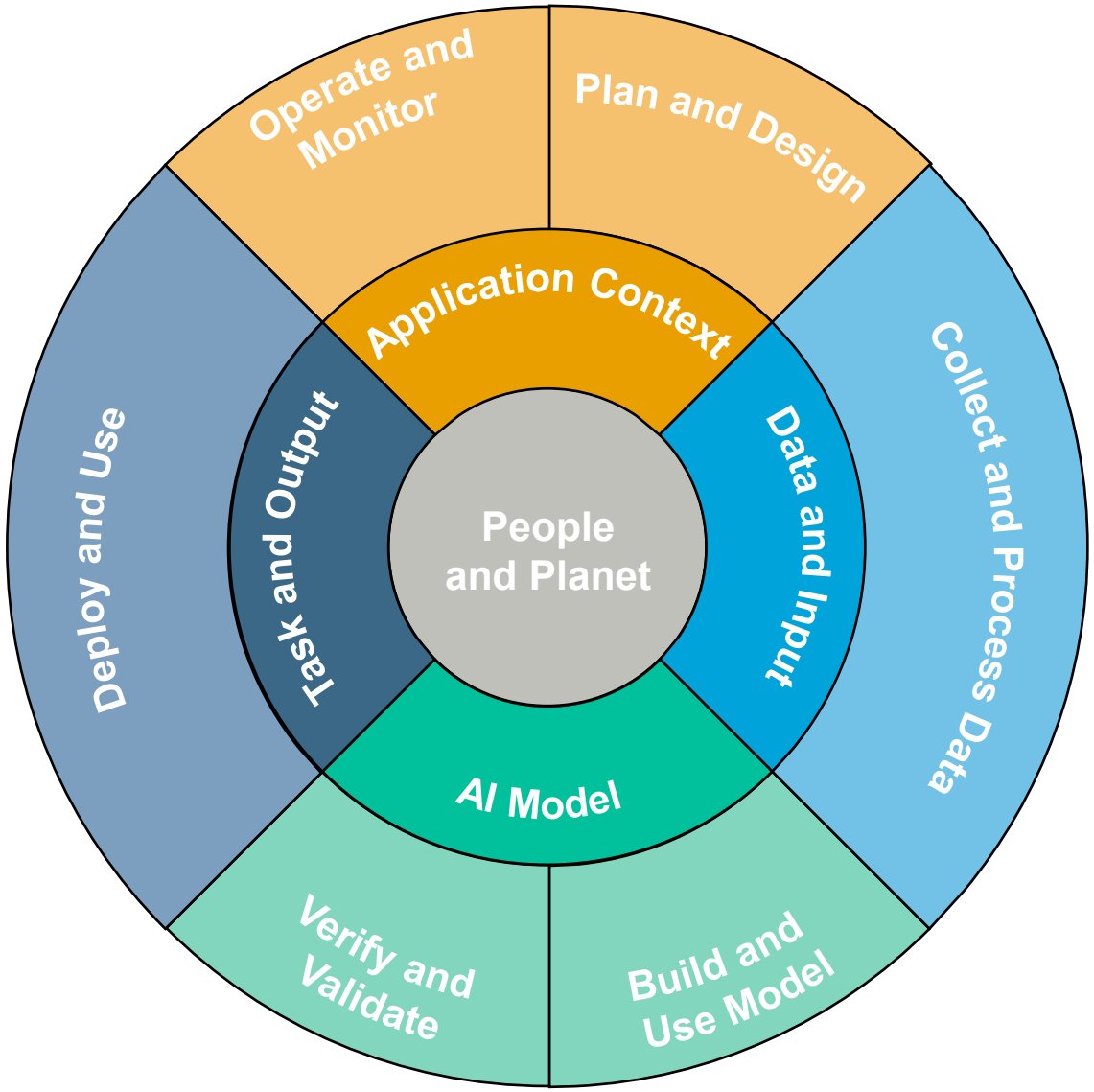
2

PRIVACY ANALYTICS
an IQVIA company

2. Guardrails Mapped to NIST AI Risk Management Framework

- The NIST AI RMF, developed in a public-private collaboration, aims to manage risks to individuals, organizations, and society that are associated with AI.
- It is organized by **Key Dimensions**, listed in the center and inner ring, and AI **Lifecycle Stages**, in the outer ring, corresponding to the use of AI tools.
- The EAB recommends guardrails for consideration to mitigate against some risks, mapped to the lower-level Lifecycle Stages of the outer ring.

For clarity, we emphasize that "Build and Use Model" in the AI RMF includes the activities of creating or selecting algorithms; training models; and model testing.



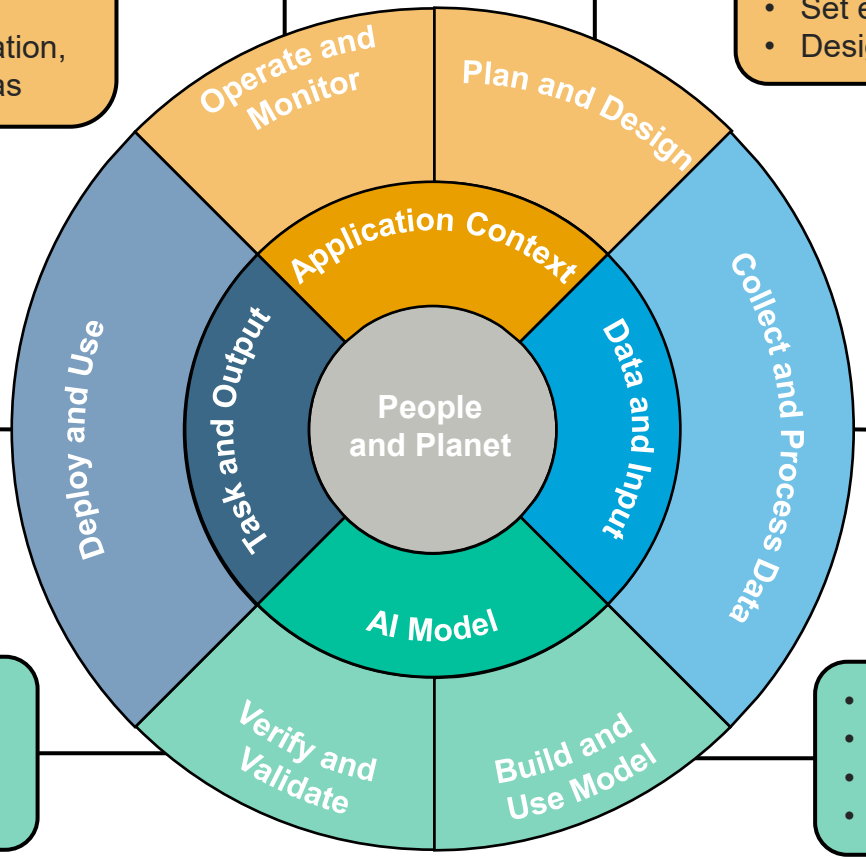
PRIVACY ANALYTICS
an IQVIA company

Framework and graphic adapted from the NIST AI Risk Management Framework. Reproduced courtesy of the National Institute of Standards and Technology, U.S. Department of Commerce. Not copyrightable in the United States

Guardrails: Safety & Adaptability

- Ensure monitoring of:
  Security, IP, identifiability, regulatory compliance, input manipulation
- Check for model skew, drift
- Check periodically for output validation, privacy/security metrics or quotas

- Establish legal basis/regulatory compliance
- Establish KPIs
- Set executive accountability and governance
- Design for system security



- Develop and share confidence scores
- Implement user-centric design, including data visualization
- Monitor for supply-chain attacks

- Track and apply data controls:
  IP, sensitivity, identifiability
- Monitor or limit prompts
- Ensure data quality
- Standardize where possible

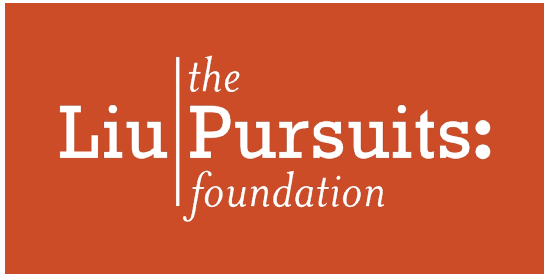
- Perform red-teaming exercises
- Check for data poisoning and input manipulation
- Validate output integrity
- Ensure results are appropriate and accurate

- Maintain intrusion protection
- Monitor for model inversion
- Mitigate against identity/membership disclosure
- Mitigate against hallucination

Safety & Adaptability
encompasses privacy and security; content review and fact checking; and adaptation and innovation

PRIVACY ANALYTICS
an IQVIA company

Example: Safety & Adaptability



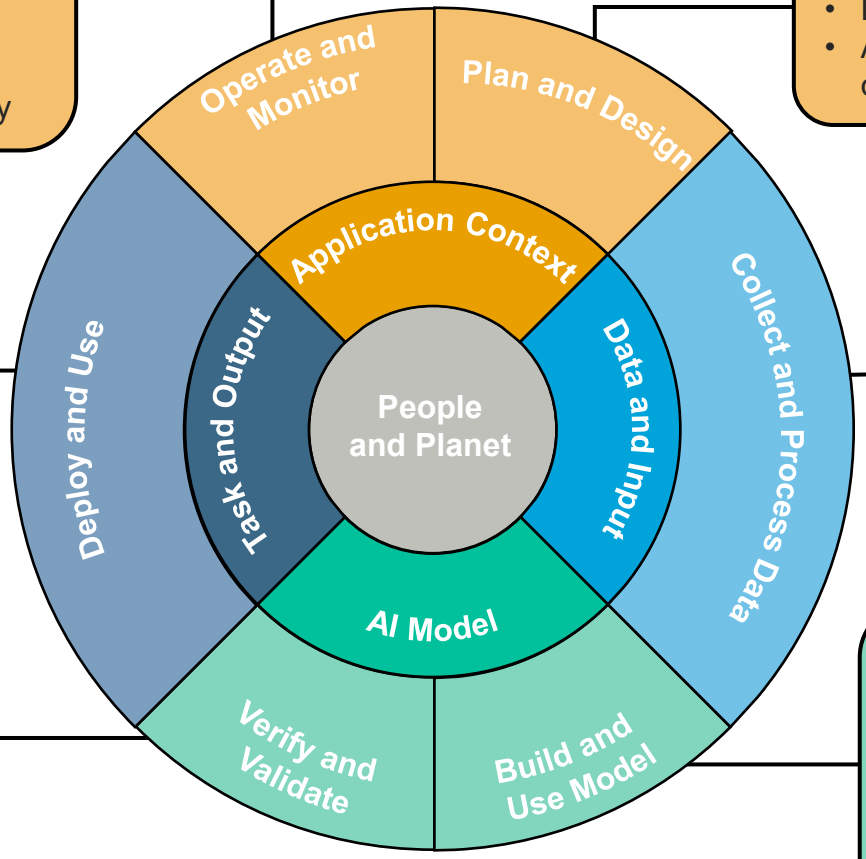
An AI driven legal guidance in SharePoint

- Setting up guardrails for how outputs may be used
- 360 input from stakeholders & iteration
- Accountability for the results

Guardrails: Governance & Transparency

- Ensure appropriate use with continuous monitoring
- Monitor for reproducibility
- Enact processes for adapting to changing standards
- Implement methods to detect misuse
- Publish outcomes to promote transparency

- Define roles and responsibilities
- Set requirements for responsible use
- Determine level of transparency
- Align to modern/best practice tech, documentation practices



- Adhere to scientific frameworks
- Provide resources:
  training data details, prompt dictionary, model output interpretation
- Evaluate impacts

- Establish accountable data owners
- Provide technical stewardship
- Implement quality standards/assurance processes
- Consider disclosure of:
  legal basis, data/model lineage, intended use cases

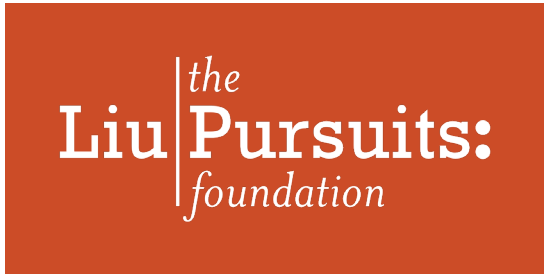
- Define and document verification and validation processes
- Benchmark against public references
- Use open source where possible
- Employ statistical validation
- Establish alerts or warnings

- Establish model accountability (model developer vs consumer)
- Seek internal and external peer review of modeling techniques
- Share KPI models
- Share statements on possibilities of model hallucinations, bias

Governance & Transparency
encompasses oversight and governance; compliance with academic and scientific standards; transparency and disclosure

PRIVACY ANALYTICS
an IQVIA company

Example: Governance & Transpacency

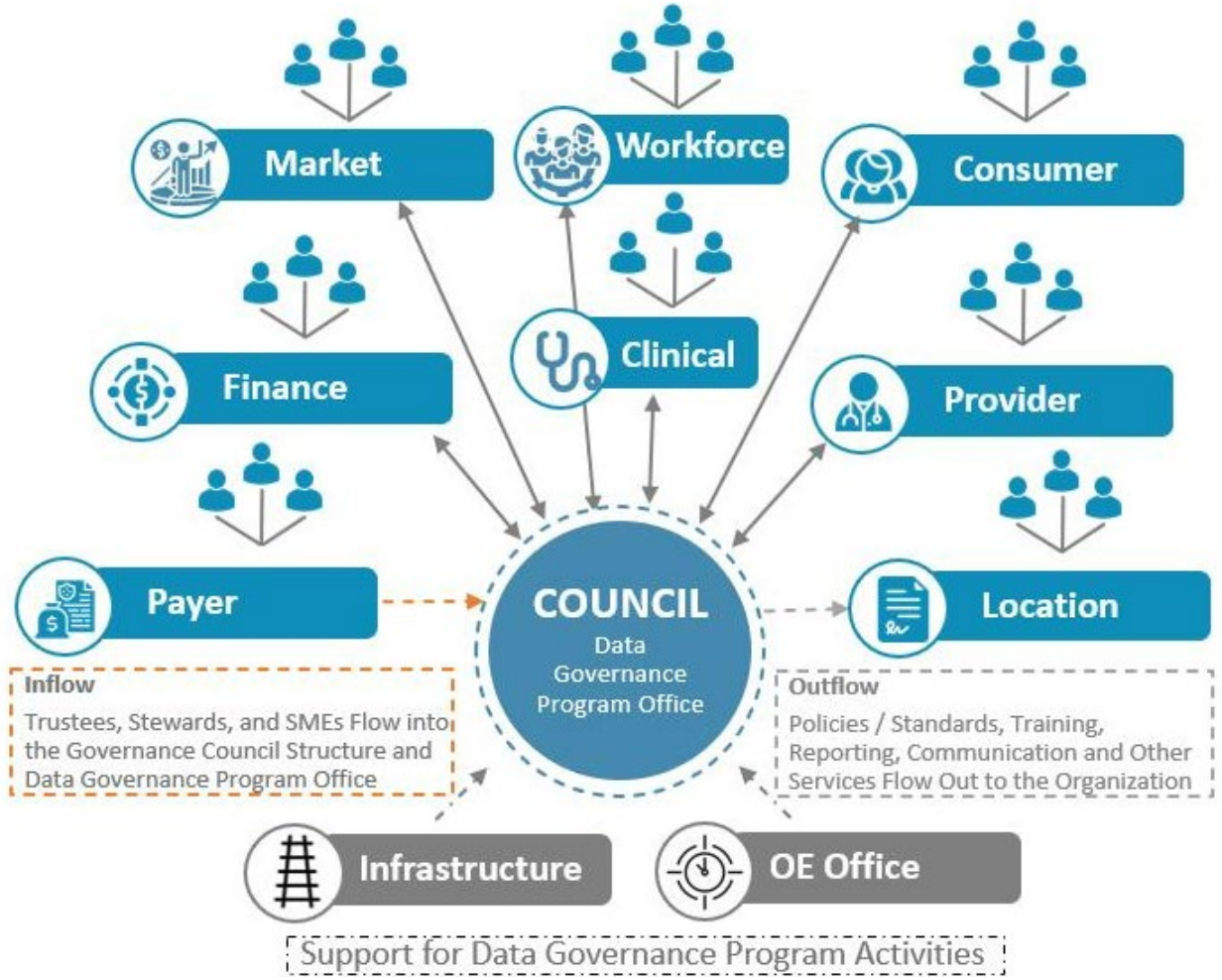


An AI driven legal guidance in SharePoint

- Periodic review of the user, validity, and value
- Providing the governance process to all users
- Ad hoc audits

Example: Governance & Transpacency

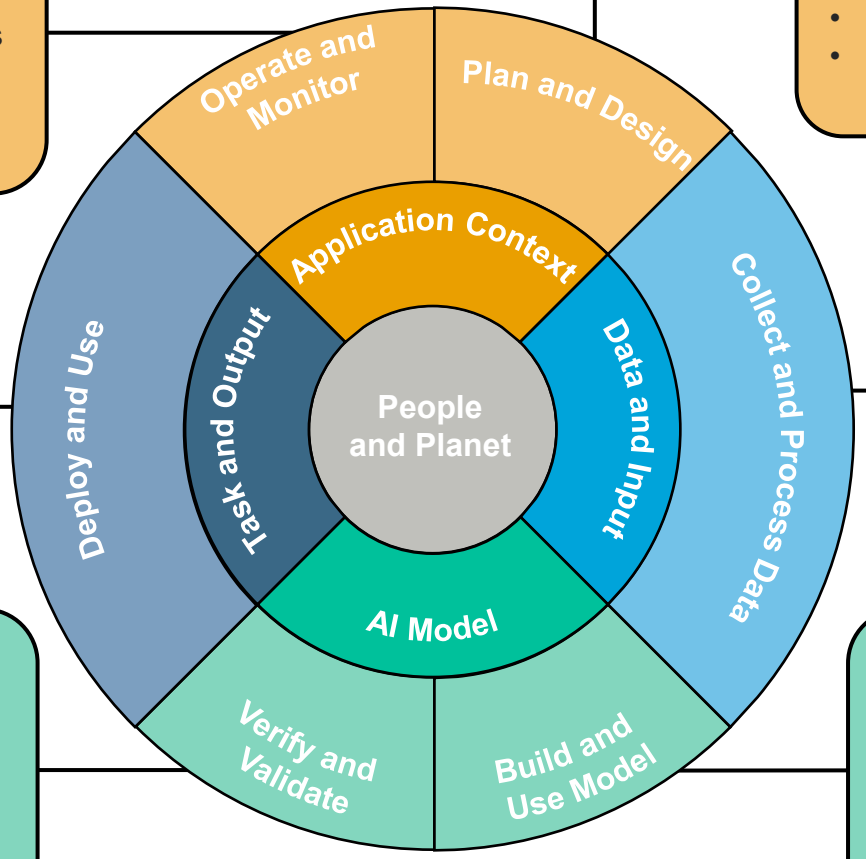
- Data Governance Framework
  - Business/Ministry Based Health Care Subject Area Domain Accountability
  - Implement new AI Domain Stewardship role
  - Align with Mercy Mission and Ethics Leadership
  - Integrate Council of Responsible AI



Guardrails: Trust & Ethical Responsibility

- Adhere to best practices
- Perform spot-checks for model drift
- Monitor for misuse
- Consider public engagement/statements on the use of AI
- Ensure ongoing informed consent with disclosures to users, regulators

- Define usage contexts
- Assess impacts of usage
- Develop process for managing bias
- Map against best practices
- Promote public engagement, informed consent/AI literacy



- Ensure consistent use context
- Maintain ongoing ethical oversight
- Apply ongoing technical robustness metrics
- Ensure user training and guidelines are available

- Verify consent, copyright
- Conduct ethical reviews of data sources and use cases
- Ensure data is representative and validate externally
- Conduct systematic checks for data bias

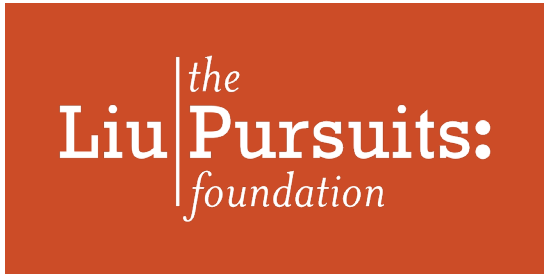
- Conduct ethics reviews of use cases, other issues
- Use human in the loop
- Check for bias in results and UX; establish clear definitions considering varied sources
- Disclose flagged good/bad prompts to model developers

- Prioritize transparency and explainability in models
- Support AI literacy across all roles
- Implement fairness metrics, balanced cohorts, other bias mitigations
- Establish prompt guardrails against inappropriate use

Trust & Ethical Responsibility
encompasses ethical and responsible uses; bias and fairness; stakeholder engagement and inclusivity

PRIVACY ANALYTICS
an IQVIA company

Example: Governance & Transpacency



An AI driven legal guidance in SharePoint

- Periodic review of the user, validity, and value
- Providing the governance process to all users
- Ad hoc audits

Example: Trust & Ethical Responsibility

- **Core Principles from Mercy's AI Governance Charter**
- **Transparency**
  - Maintain openness about AI development, deployment, and usage.
  - Ensure stakeholders understand how AI decisions are made.
- **Accountability**
  - Establish clear lines of responsibility for AI decisions and outcomes.
  - Humans remain in control and accountable for AI systems.
- **Ethics and Compliance**
  - Align AI use with organizational values and ethical standards.
  - Comply with all relevant laws and regulations, including privacy and security requirements.
- **Fairness and Non-Discrimination**
  - Prevent bias and discriminatory outcomes in AI applications.
  - Promote equitable treatment across all user groups.
- **Privacy and Security**
  - Protect individual rights and secure all data used in AI initiatives.
  - Implement safeguards against unauthorized access and misuse.
- **Human-Centric and Socially Beneficial**
  - Keep AI applications focused on improving patient care and societal well-being.
  - Ensure technology serves people, not replaces human judgment.
- **Continuous Monitoring and Improvement**
  - Regularly audit AI models for performance, fairness, and compliance.
  - Update practices as technology and regulations evolve.
- **Collaboration and Education**
  - Foster cross-functional teamwork and knowledge sharing.
  - Invest in training to build AI literacy across the organization

Thank you!

Jean Liu
Executive Director,
Liu Pursuits

Doug Graham
Director of Enterprise
Data Governance,
Mercy

Ren-Yi Lo
Head of Autonomous
Sys & Data Gov,
Siemens Healthineers,
AI Tech Center

Brian Rasquinha
Associate Director,
Solution Architecture,
Privacy Analytics

PRIVACY ANALYTICS
an IQVIA company