Guardrails for Generative Al as Part of a Safe & Scalable Al Program

Executive Advisory Board

Version 1 - September 4, 2025



Introduction

- The Executive Advisory Board (EAB) is comprised of 21 senior-level executives from public and private sector organizations. EAB members are based in North America and the UK/EU, and primarily from healthcare, life sciences, and financial services contexts. EAB members participate as individuals, and their views do not necessarily reflect that of their respective organizations.
- The EAB serves as a forum for the exchange of insights on key industry issues pertaining to the safe and
 responsible use of sensitive information to drive innovation. The EAB aims to produce public-facing thought
 leadership to benefit the broader community of professionals working to safely release value from data
 derived from information about people, and help organizations earn trust. The EAB is facilitated by Privacy
 Analytics, an IQVIA company.
- The EAB identified a need for additional guidance for evaluating use cases and establishing appropriate guardrails as organizations move to increase their use of artificial intelligence (AI) technologies, in particular Generative AI (GenAI) in response to market demands. The additional guidance is built on a core of the NIST AI Risk Management Framework. The EAB additionally acknowledges other existing guidance:
 - 5 Safes Framework, applied to data de-identification and anonymization
 - Coalition for Healthcare AI (CHAI)'s Responsible AI Guide (RAIG) and Checklists (RAIC)



Introduction

- The EAB has created this guidance with the purpose of empowering leaders in data, analytics, privacy, legal, and AI with a tool to help achieve alignment on the use of GenAI within their organizations. Emphasizing safe and responsible use of sensitive information and emerging technology can drive public and partner trust while supporting organizations in building efficiency and responding to opportunities for innovation.
- The guidance is comprised of two components:
 - **Risk-level factors**. This component can provide a basis by which an organization can qualitatively evaluate the risk level of a use case and thus determine a relative level of guardrails the use-case may merit.
 - **Suggested guardrails**. This component provides guardrails, across all stages of the Al lifecycle, for consideration when implementing the Al initiative.



Introduction

- The guidance is 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.
- Guardrails are intended to contribute to risk mitigation, which is intended to mean reasonable and balanced reduction of risks rather than, for example, zero risk. This guidance does not claim to result in minimal achievable risk, nor does it imply any persistent residual risk is unacceptable.
- The guidance is 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.
- Al risks are contextually dependent. Risk assessments may vary depending on the sector, application, institution, or other domain-specific factors.



1. Evaluating Risk Level for Use Cases across Multiple Dimensions

- This first tool lists dimensions that can be considered as affecting the **organizational risk** associated with the use of generative AI. These dimensions are listed as separate rows on the leftmost side in the figure on the following graphic.
- Features or factors are presented across a spectrum from "Lower Risk" to "Higher Risk" to illustrate the general impact on organizational risk.
- All dimensions (or rows) are intended to be taken into consideration. It is possible a use case may have a
 mix of low and high-risk dimensions. The overall risk level is intended to reflect a consolidated posture, and
 may err on conservatively consolidating towards the higher end of risk.
- Use cases tending towards Lower Risk or Higher Risk responses may be treated accordingly, with higherrisk use cases indicating an increased need for stronger or more numerous guardrails.
 - Guardrails will be described in the following tool.



1. Evaluating Risk Level for Use Cases across Multiple Dimensions

	Lower Risk			Higher Risk
Context of Users	Under Internal to contract 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



2. Guardrails Mapped to NIST AI Risk Management Framework

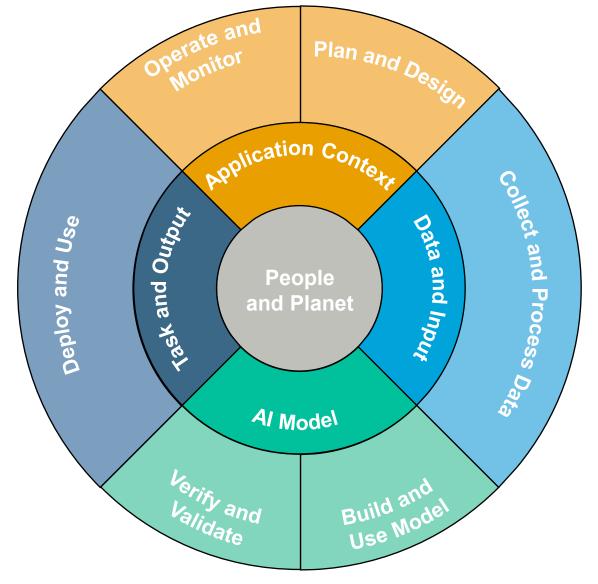
- This second tool provides a set of guardrails considered by the EAB.
- The guardrails are mapped to the structure of the <u>NIST AI Risk Management Framework</u> (AI RMF), which in turn is aligned to the AI lifecycle.
- The guardrails are further broken down into three categories of approaches:
 - Safety & Adaptability
 - Governance & Transparency
 - Trust & Ethical Responsibility
- Use cases with higher organizational risk would typically benefit from stronger, more numerous, and/or broader guardrails for the use of AI.



2. Guardrails Mapped to NIST Al Risk Management Framework

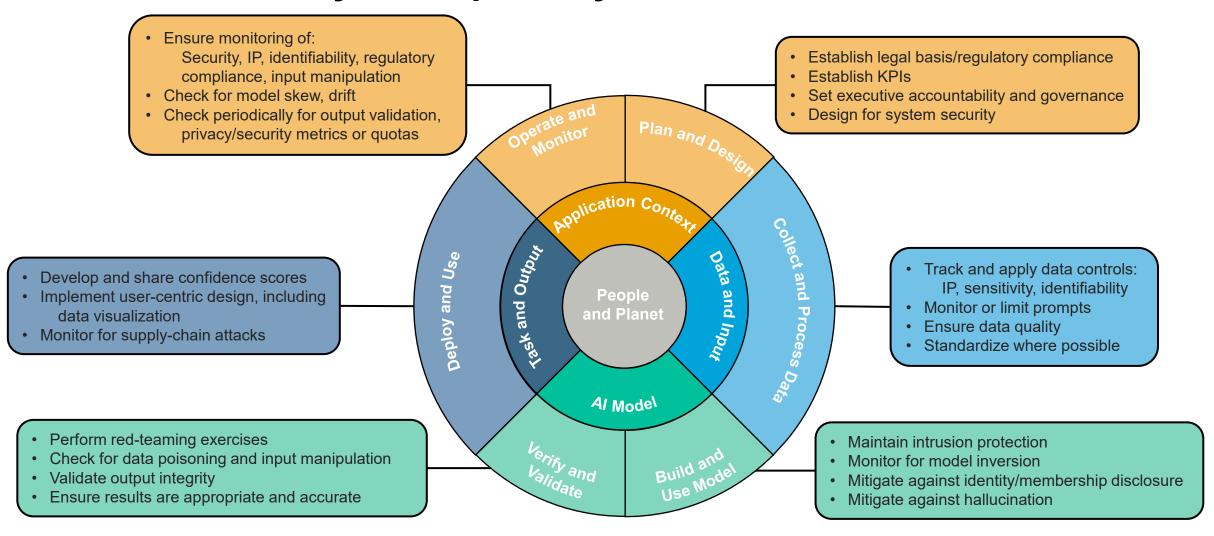
- The NIST AI RMF, developed in a publicprivate collaboration, aims to manage risks to individuals, organizations, and society that are associated with AI.
- It is organized by Key Dimensions, listed in the center and inner ring, and Al Lifecycle Stages, in the outer ring, corresponding to the use of Al 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.





Guardrails: Safety & Adaptability





Safety & Adaptability

encompasses privacy and security; content review and fact checking; and adaptation and innovation

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

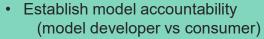
- Adhere to scientific frameworks
- Provide resources: training data details, prompt dictionary, model output interpretation
- Evaluate impacts

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

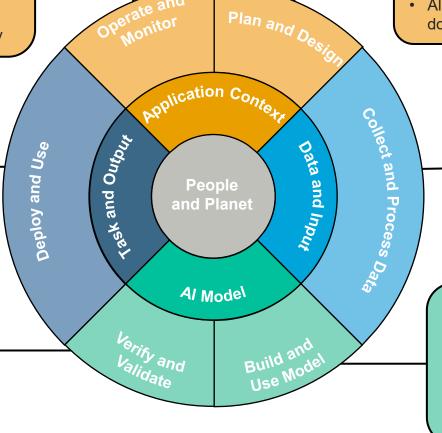
- Define roles and responsibilities
- · Set requirements for responsible use
- Determine level of transparency
- Align to modern/best practice tech, documentation practices



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



- Seek internal and external peer review of modeling techniques
- Share KPI models
- Share statements on possibilities of model hallucinations, bias



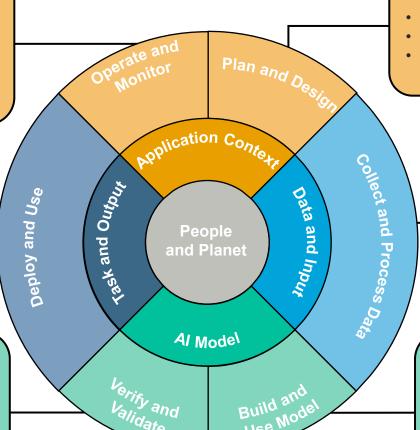


Governance & Transparency

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

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



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

- 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
- Prioritize transparency and explainability in models
- Support Al 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

Contributing EAB Members

• This guidance was developed by the Executive Advisory Board (EAB), focused on supporting safe and scalable Al programs. The EAB is comprised of 21 senior-level executives working in data, analytics, privacy, and legal roles in private and public organizations. The following people are some of the EAB members who contributed to this guidance:

Paul White, SVP, Data Insights, Finthrive

Jenn Geetter, Partner, MWS

Courtney Bowman, Global Director, Privacy & Civil Liberties, Palantir

Doug Graham, Director of Enterprise Data Governance, Mercy

Ren-Yi Lo, Head of Autonomous Systems & Data Governance, Siemens Healthineers, Al Tech Center

Chris Allison, Director General, Data Analytics & Information Management, Department of National Defense (Canada) **Marlon Domingus**, Data Protection Officer in Residence, Erasmus University Rotterdam

Jean Liu, Executive Director, Liu Pursuits

Gabriel S. Eichler, PhD, Managing Director, Oak Health Partners AG

Kelly Ko, VP, Innovation, Banner Health

Phil Lindemann, VP, Data and Research, Epic

The EAB is supported and facilitated by Privacy Analytics, an IQVIA company.

- Jordan Collins, General Manager of Privacy Analytics & EAB Executive Sponsor
- Luk Arbuckle, Global Al Practice Leader, IQVIA & EAB Topic Facilitator
- Santa Borel, Associate Director, Data Privacy Solutions & EAB Topic Facilitator
- Brian Rasquinha, Associate Director, Solution Architecture & EAB Program Director
- Graham Machacek, Director, Strategy & Marketing, IQVIA & EAB Advisor

For more information, please contact

Brian Rasquinha at brasquinha@privacy-analytics.com or Jordan Collins at jocalins@privacy-analytics.com

