Day 3

Generative AI Security

Risk-Based Approach to AI Governance:

What Is an AI Risk?

An AI risk is any possibility that an AI system may cause harm, fail, or create negative impact for users, business, or society. Simply, it talks about *"What bad thing can happen because of this AI system?"*

Categories of AI risks: (Generated using Gemini)

Risk Assessment Techniques:

AI risk assessment evaluates how dangerous an AI system could be. Risk assessment can be qualitative (Uses words, not numbers), quantitative (Uses numbers, scoring, metrics, and probabilities), or scoring-based.

Risk Prioritization & Heat Maps:

Heat maps help prioritize which risks need immediate attention. After, scoring risks, we organize them in a heat map:

Likelihood	Impact	Risk Level
High + High	🔥 Red	Critical
High + Medium	🟧 Orange	High
Medium + Medium	🟨 Yellow	Medium
Low + Low	🟩 Green	Low

Simply,

Red risks → fix immediately
Green risks → normal monitoring
Yellow → moderate controls

Linking Risks to Controls and Mitigations:

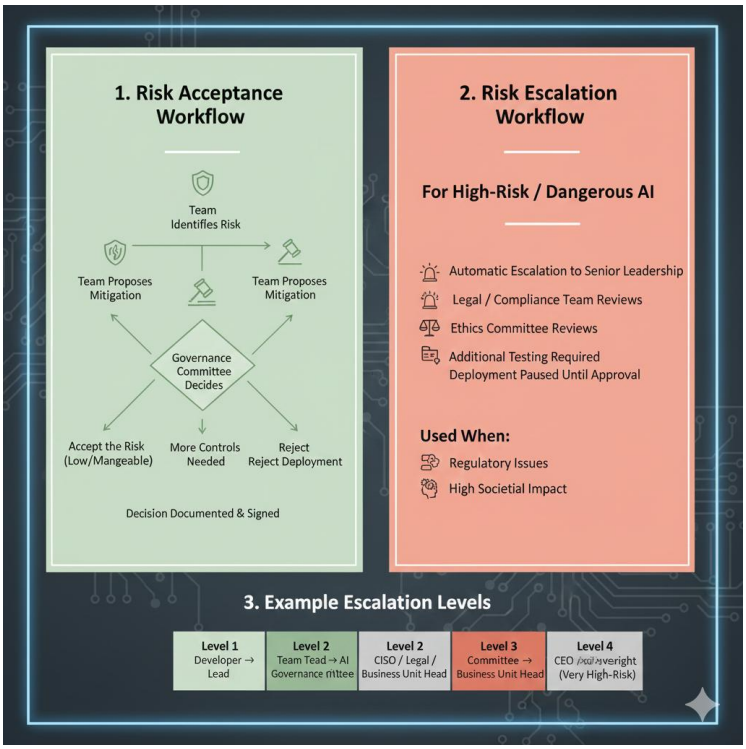
For every risk identified, governance requires specific controls (A control is a safeguard, procedure, or mechanism put in place to reduce, prevent, or manage a specific risk.). Governance ensures every risk has a matching control.

Example Mapping:

Risk	Control / Mitigation
Biased data	Fairness testing, rebalancing datasets
Model drift	Continuous monitoring, retraining
Prompt injection	Strong input validation, guardrails
Model theft	API rate limits, access control, encryption
Privacy leakage	Differential privacy, anonymization
Non-compliance	Documentation, human oversight, audit logs

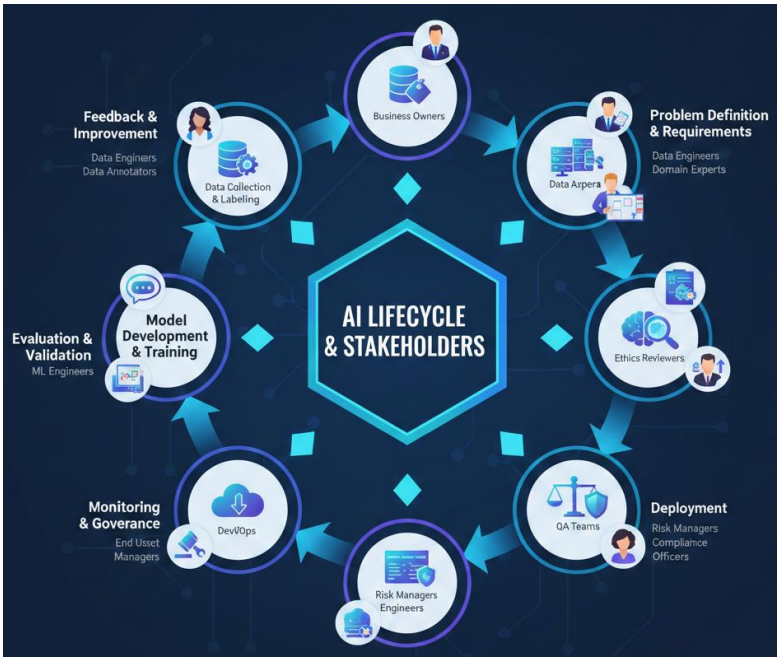
Governance Workflows for Risk Acceptance & Escalation:

If a risk cannot be fully removed, governance defines what to do. Governance workflows define who can accept, reject, or escalate high-risk AI decisions.



AI Stakeholders & Their Roles:

AI lifecycle and stakeholders:



- Many people work together to make AI safe.
- ML engineers build models, data teams prepare data, security protects systems.
- Compliance ensures laws are followed, business owners justify use cases.
- Executives give final approval, users provide feedback.

RACI Models for AI Projects:

RACI = Responsible, Accountable, Consulted, Informed

Typical matrix:

Activity	Responsible	Accountable	Consulted	Informed
Data collection	Data team	Data owner	Legal/Compliance	Executives
Model training	ML Engineers	AI Lead	Security, Data	Business owner
Security testing	Security team	CISO	ML Engineers	Executives
Risk assessment	Compliance	CISO/CTO	ML + Security	Business owner
Deployment	ML Ops	CTO	Security + Compliance	All stakeholders
Monitoring	Security + ML Ops	CISO	Business	Regulators (if needed)

RACI helps avoid confusion about who does what.

Note: Collaboration happens during training, deployment, and monitoring.

--The End--