Suganya Purushothaman

AI Product Manager | Governance, Security & Responsible AI

AI/ML Security & Governance

General Principles

The following principles apply across all AI/ML domains, ensuring that models, systems, and user experiences are designed with governance, trust, and compliance at their core.

- Data Minimization & PII Protection
- Access Control & Immutable Audit Logs
- Secure Inference & Encryption
- Bias, Fairness, and Transparency
- Governance Gates before release
- Monitoring, Drift Detection, and Retraining
- Human-in-the-loop for sensitive use cases
- Vendor and Third-party risk assessments

Governance in Recommendation Systems

Recommendation systems, especially in sensitive domains like children's education or healthcare, require additional governance to ensure trust and fairness:

- Dataset curation: Regularly audit datasets to avoid bias (e.g., over-representing certain genres or demographics).
- Explainability: Provide parents or users with 'Why was this recommended?' explanations.
- Personalization Privacy: Ensure consent when collecting age, interest, or behavioral data.
- Evaluation: Use precision, recall, and diversity metrics to ensure fair and balanced recommendations.

Governance in AI Dubbing Systems

AI-based dubbing systems introduce risks around voice data, translations, and cultural context. To ensure safe adoption:

- Voice Data Privacy: Avoid storing raw child/parent voices; anonymize training data.
- Translation Safety: Vet translations for cultural sensitivity and avoid offensive language.
- Model Security: Prevent unauthorized voice cloning or misuse.
- Compliance: Ensure use of dubbing for children's media aligns with COPPA, GDPR, and local DPDP regulations.

Governance in Chatbots

Chatbots designed for wellness, parenting, or sensitive conversations must have built-in safeguards:

- Guardrails: Use retrieval firewalls to block unsafe queries and inject safe defaults.
- Tone & Empathy: Fine-tune models to respond in safe, supportive, and culturally aware ways.
- Escalation Paths: Route high-risk queries to human counselors or verified professionals.
- Logging & Transparency: Maintain clear logs of chatbot conversations for audit without storing unnecessary PII.

Cross-Domain Governance Extensions

Across recommendation systems, dubbing AI, and chatbots, governance must be embedded into the system lifecycle:

- Multi-layer firewalls (input filtering, prompt injection detection, output moderation).
- Cost vs. Safety Trade-offs: Use caching, batching, and fallback layers to balance efficiency with compliance.
- Observability: Track metrics not only for performance but also for fairness, bias, and safety events.
- Regular Audits: Independent governance reviews every quarter for compliance and risk assessment.

AI/ML Security & Governance — 10 Practical Steps

1. Data Minimization Collect and store only the minimum data necessary for the model. Remove or mask personally identifiable information (PII) before ingestion. This reduces privacy risks and compliance overhead.

2. Access Control & Audit Logs Restrict access to sensitive datasets and model endpoints. Maintain immutable audit logs to track who accessed what and when, supporting accountability and investigations.

3. PII Filtering & Anonymization Before sending data to external APIs or cloud services, apply PII filters and anonymization. Replace sensitive tokens with pseudonyms to safeguard user identities.

4. Model Explainability Document why a model was chosen, its inputs, and its decision-making process. Provide simple explainability notes for business stakeholders to build trust.

5. Secure Inference Ensure inference requests are encrypted in transit (HTTPS/TLS). Use private endpoints, authentication, and rate-limiting to prevent misuse of your models.

6. Bias & Fairness Checks Continuously test datasets and model outputs for bias. Track fairness metrics and publish mitigation plans as part of release notes to ensure accountability.

7. Governance Gate for Releases Introduce a formal release checklist. Each model version must pass governance checks: data quality, bias, privacy, reproducibility, and cost thresholds.

8. Human-in-the-loop & Escalation For high-risk use cases (e.g., healthcare, finance), involve human reviewers before final decisions. Define clear escalation workflows for sensitive outputs.

9. Monitoring & Drift Detection Deploy monitoring for input distribution shifts and performance decay. Trigger retraining, alerts, or rollbacks when drift or anomalies are detected.

10. Contracts & Third-party Assessment When integrating external APIs or third-party models, vet them for compliance and add clear SLA/security clauses. Ensure vendors meet your governance standards.