

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.