Governing the Machines: Executable AI Governance in the Agentic Era

Guest Lecture by **Adnan Masood, PhD**

Graduate Seminar

Fall 2025

Agenda

1. Taxonomy: the AI Governance Stack
2. Concepts: term $\rightarrow$ definition $\rightarrow$ intuition (interactive)
3. Quantitative fairness & bias metrics (with formulas)
4. Why it matters: five real incidents (+ what went wrong)
5. Controls: observability, security, responsible AI
6. Agentic AI: why autonomy amplifies risk
7. Industry use cases & regulatory overlays
8. Key regulations & timelines (EU, US federal, US states, standards)
9. What it means for you (executive, engineer, manager, IC)
10. Team exercise: use case $\rightarrow$ regulation $\rightarrow$ controls
11. Summary, references, and appendix checklists

Taxonomy of Ideas

The AI Governance Stack (at a glance)

Strategy & Principles

- Ethical / Responsible AI
- Risk appetite & tolerances

Governance

- Policies & standards
- Roles (Board, GC, CISO, CDO, Eng)

Controls & Assurance

- Preventive, detective, corrective controls
- Independent model validation & audit

Operations & Oversight

- ModelOps / AI TRiSM cadence
- Observability & incident response

Compliance & Regulation

- Laws, rules, guidances
- Certifications & attestations

Outcomes

- Safety, Security, Fairness, Privacy
- Transparency, Accountability, Reliability

Concepts (Interactive Triads)

Governance

Governance — Definition

Decision rights, policies, standards, and checkpoints that control how AI is built, deployed, and used across the organization.

Governance — Intuition & Example

Intuition. Think of it as the corporate constitution for AI—the who/what/when of decision-making.

Example. Board-approved AI policy, RACI for model lifecycle, and go/no-go gates before release.

Compliance

Compliance — Definition

Producing evidence that systems meet applicable laws, regulations, and internal policies.

Compliance — Intuition & Example

Intuition. If it's not evidenced, it didn't happen.

Example. Evidence packs: model cards, audit logs, DPIAs/PIAs, evaluation results, third-party attestations.

Regulation

Regulation — Definition

Binding rules issued by governments or regulators (with penalties for non-compliance).

Regulation — Intuition & Example

Intuition. Defines the minimum operating baseline; often technology-agnostic but outcome-focused.

Example. EU AI Act GPAI obligations; NYC LL 144 bias audit; California CPPA ADMT rights/assessments.

Audit

Audit — Definition

Independent examination to verify controls exist and operate effectively.

Audit — Intuition & Example

Intuition. Trust, but verify—by someone who does not run the system.

Example. Internal Audit or third-party performs model risk audit against policy & ISO 42001.

Control

Control — Definition

A specific mechanism (process or technical) designed to reduce a defined risk.

Control — Intuition & Example

Intuition. Guardrails that prevent, detect, or correct failure modes.

Example. Preventive: role-based access. Detective: bias dashboards. Corrective: kill-switch/rollback.

Assurance

Assurance — Definition

Confidence backed by evidence (audits, certifications, attestations) that risks are managed.

Assurance — Intuition & Example

Intuition. Show, don't tell.

Example. SOC 2, ISO 27001/42001 certificates; published bias audit summaries.

Risk

Risk — Definition

Possibility of harm or loss from AI-driven actions, measured as likelihood $\times$ impact.

Risk — Intuition & Example

Intuition. Risk increases with autonomy, reach, and coupling to real-world tools.

Example. Discrimination, data leakage, unsafe outputs, security compromise, operational failures.

Ethical AI

Ethical AI — Definition

Value-aligned choices that avoid foreseeable harm and respect rights and dignity.

Ethical AI — Intuition & Example

Intuition. The north star: do no harm and respect people as ends, not means.

Example. Ban manipulative uses; require transparency for chatbots; human appeal paths.

Responsible AI

Responsible AI — Definition

Operationalizing ethics via policy, controls, testing, and accountability.

Responsible AI — Intuition & Example

Intuition. Make ethics executable.

Example. Policy-to-control mapping; thresholds; red-teaming; signoffs; continuous monitoring.

Fairness

Fairness — Definition

Outcomes that do not systematically disadvantage protected groups.

Fairness — Intuition & Example

Intuition. Comparable individuals should face comparable error rates/opportunities.

Example. Set quantitative thresholds for TPR/FPR gaps and disparate impact ratio.

Bias

Bias — Definition

Systematic error from data, model, or process that skews outcomes.

Bias — Intuition & Example

Intuition. Bias can enter at data generation, labeling, modeling, evaluation, or deployment.

Example. Historical, representation, measurement, aggregation, evaluation biases.

Transparency

Transparency — Definition

Disclosing how/why an AI system works and is used.

Transparency — Intuition & Example

Intuition. No black boxes when stakes are high.

Example. Plain-language model cards; data lineage; change logs; user notices.

Explainability

Explainability — Definition

Making model decisions understandable to humans (local or global).

Explainability — Intuition & Example

Intuition. Right level of detail for the audience and risk.

Example. Feature attributions, counterfactual explanations, examples of similar cases.

Safety

Safety — Definition

Preventing physical, psychological, or operational harm from AI behavior.

Safety — Intuition & Example

Intuition. Avoids catastrophic or high-severity failure modes.

Example. Safety cases; red-team stress tests; incident drills; safe defaults and fallbacks.

Security

Security — Definition

Protecting the AI system, its data, and its tools from adversaries.

Security — Intuition & Example

Intuition. Assume attackers will attempt prompt injection, data poisoning, model theft.

Example. Isolation of tools/retrievers, content filters, rate limits, anomaly detection.

Privacy

Privacy — Definition

Limiting collection, processing, and sharing to lawful, necessary, and proportionate use.

Privacy — Intuition & Example

Intuition. Data minimization is the cheapest control you have.

Example. DPIAs/PIAs; purpose limitation; deletion SLAs; synthetic or federated options.

Observability

Observability — Definition

End-to-end visibility of data, models, and decisions with traceable logs.

Observability — Intuition & Example

Intuition. If you can't see it, you can't fix it.

Example. Decision logs, model cards, lineage, eval dashboards, drift/bias monitors, alerts.

Model Risk Management (MRM)

Model Risk Management (MRM) — Definition

Structured lifecycle to identify, measure, mitigate, and monitor model risk.

Model Risk Management (MRM) — Intuition & Example

Intuition. Three lines of defense: builders, validators, auditors.

Example. Design review, independent validation, periodic revalidation, change control.

Agentic AI

Agentic AI — Definition

AI systems that plan, call tools, and act autonomously toward goals.

Agentic AI — Intuition & Example

Intuition. Autonomy compounds blast radius and couples AI errors to real-world actions.

Example. Agents issuing payments, emails, code commits, or ticket updates with minimal oversight.

Prompt Injection

Prompt Injection — Definition

Inputs crafted to hijack model behavior, instructions, or tools.

Prompt Injection — Intuition & Example

Intuition. Treat inputs like untrusted code.

Example. User content instructs the agent to exfiltrate secrets or ignore policies.

Hallucination

Hallucination — Definition

Confidently stated output that is false or ungrounded.

Hallucination — Intuition & Example

Intuition. In open-ended tasks, models will 'complete' patterns even when data is missing.

Example. Fabricated citations; invented API parameters; non-existent legal precedents.

Data Drift

Data Drift — Definition

Change in data distribution between training and production (covariate, prior, or concept).

Data Drift — Intuition & Example

Intuition. Your model is only as good as yesterday's data.

Example. PSI/KL divergence alerts; re-training triggers; canaries; shadow deployments.

Red Teaming

Red Teaming — Definition

Systematic adversarial testing to discover failure modes in safety & security.

Red Teaming — Intuition & Example

Intuition. Attack yourself before attackers do.

Example. Jailbreak suites; tool-abuse scenarios; RAG data exfiltration; agent loop tests.

Human-in-the-Loop

Human-in-the-Loop — Definition

Humans with authority to approve, override, or appeal AI decisions.

Human-in-the-Loop — Intuition & Example

Intuition. Humans as circuit-breakers, not rubber stamps.

Example. Manual review thresholds; dual control on sensitive actions; appeal portals.

Quantitative Fairness & Bias Metrics

Demographic Parity (Statistical Parity)

Demographic Parity (Statistical Parity) (Metric)

Definition. Positive decision rates are independent of protected attribute.

Formula.

$$P(\widehat{Y} = 1 \mid A = a) = P(\widehat{Y} = 1 \mid A = b) \quad \forall a, b$$

Notes. Often unrealistic in safety-critical domains; can mask qualification differences.

Demographic Parity (Statistical Parity) — Intuition & Example

Intuition. When base rates differ, trade-offs emerge; pick a metric aligned to harms.

Example. Use domain review to set thresholds; monitor gaps over time with CIs.

Disparate Impact Ratio (80% Rule)

Disparate Impact Ratio (80% Rule) (Metric)

Definition. Selection rate of protected group should be at least 80% of the highest group's rate.

Formula.

$$\mathrm{DIR} = \frac{P(\widehat{Y} = 1 \mid A = \text{protected})}{P(\widehat{Y} = 1 \mid A = \text{reference})} \geq 0.8$$

Notes. EEOC four-fifths rule; initial screen for adverse impact (not dispositive).

Disparate Impact Ratio (80% Rule) — Intuition & Example

Intuition. When base rates differ, trade-offs emerge; pick a metric aligned to harms.

Example. Use domain review to set thresholds; monitor gaps over time with CIs.

Equal Opportunity

Equal Opportunity (Metric)

Definition. True positive rates (TPR/recall) are equal across groups.

Formula.

$$\mathrm{TPR}_a = \mathrm{TPR}_b \;\Rightarrow\; P(\widehat{Y} = 1 \mid Y = 1, A = a) = P(\widehat{Y} = 1 \mid Y = 1, A = b)$$

Notes. Focuses on qualified individuals receiving positive outcomes.

Equal Opportunity — Intuition & Example

Intuition. When base rates differ, trade-offs emerge; pick a metric aligned to harms.

Example. Use domain review to set thresholds; monitor gaps over time with CIs.

Equalized Odds

Equalized Odds (Metric)

Definition. Both TPR and FPR are equal across groups.

Formula.

$$\mathrm{TPR}_a = \mathrm{TPR}_b \ \wedge \ \mathrm{FPR}_a = \mathrm{FPR}_b$$

Notes. Stronger constraint; may reduce accuracy; requires threshold or post-processing.

Equalized Odds — Intuition & Example

Intuition. When base rates differ, trade-offs emerge; pick a metric aligned to harms.

Example. Use domain review to set thresholds; monitor gaps over time with CIs.

Predictive Parity

Predictive Parity (Metric)

Definition. Positive predictive value (precision) is equal across groups.

Formula.

$$\mathrm{PPV}_a = \mathrm{PPV}_b \;\Rightarrow\; P(Y = 1 \mid \widehat{Y} = 1, A = a) = P(Y = 1 \mid \widehat{Y} = 1, A = b)$$

Notes. In tension with equalized odds when base rates differ.

Predictive Parity — Intuition & Example

Intuition. When base rates differ, trade-offs emerge; pick a metric aligned to harms.

Example. Use domain review to set thresholds; monitor gaps over time with CIs.

Calibration Within Groups

Calibration Within Groups (Metric)

Definition. Predicted probability equals observed outcome rate within each group.

Formula.

$$P(Y = 1 \mid \widehat{P} = p, A = a) \approx p \quad \forall p, a$$

Notes. Ensures scores have the same meaning across groups.

Calibration Within Groups — Intuition & Example

Intuition. When base rates differ, trade-offs emerge; pick a metric aligned to harms.

Example. Use domain review to set thresholds; monitor gaps over time with CIs.

Average Odds Difference

Average Odds Difference (Metric)

Definition. Average difference in TPR and FPR across groups.

Formula.

$$\mathrm{AOD} = \tfrac{1}{2}\left[(\mathrm{FPR}_a - \mathrm{FPR}_b) + (\mathrm{TPR}_a - \mathrm{TPR}_b)\right]$$

Notes. Zero is ideal; complements DIR/SPD.

Average Odds Difference — Intuition & Example

Intuition. When base rates differ, trade-offs emerge; pick a metric aligned to harms.

Example. Use domain review to set thresholds; monitor gaps over time with CIs.

Statistical Parity Difference

Statistical Parity Difference (Metric)

Definition. Difference in positive rates across groups.

Formula.

$$\mathrm{SPD} = P(\widehat{Y} = 1 \mid A = a) - P(\widehat{Y} = 1 \mid A = b)$$

Notes. Close to zero is desired; direction indicates advantaged group.

Statistical Parity Difference — Intuition & Example

Intuition. When base rates differ, trade-offs emerge; pick a metric aligned to harms.

Example. Use domain review to set thresholds; monitor gaps over time with CIs.

Treatment Equality

Treatment Equality (Metric)

Definition. Ratio of FN to FP is equal across groups.

Formula.

$$\frac{\mathrm{FN}_a}{\mathrm{FP}_a} = \frac{\mathrm{FN}_b}{\mathrm{FP}_b}$$

Notes. Balances over- and under-enforcement impacts.

Treatment Equality — Intuition & Example

Intuition. When base rates differ, trade-offs emerge; pick a metric aligned to harms.

Example. Use domain review to set thresholds; monitor gaps over time with CIs.

Counterfactual Fairness

Counterfactual Fairness (Metric)

Definition. A prediction is fair if it is invariant under counterfactual changes to protected attributes.

Formula.

$$\widehat{Y}_{A \leftarrow a}(U) = \widehat{Y}_{A \leftarrow a'}(U) \quad \forall a, a'$$

Notes. Requires causal model of the data-generating process.

Counterfactual Fairness — Intuition & Example

Intuition. When base rates differ, trade-offs emerge; pick a metric aligned to harms.

Example. Use domain review to set thresholds; monitor gaps over time with CIs.

Why It Matters: Real Incidents

Air Canada chatbot liability

Air Canada chatbot liability — One-line story

A court held the airline liable for misinformation given by its website chatbot to a grieving customer seeking bereavement fares.

Source:
https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-february/bc-tribunal-confirms-companies-remain-liable-information-provided-ai-chatbot/

Air Canada chatbot liability — What went wrong & fixes

- What went wrong: No human-in-the-loop for sensitive policies; weak disclosures; inadequate oversight and logs. **Remedy:** governance for consumer-facing bots, approval workflows, and published policy sources.

Cruise robotaxi pedestrian dragging

Cruise robotaxi pedestrian dragging — One-line story

An autonomous vehicle struck and dragged a pedestrian; reporting omissions and safety concerns led to fines, license actions, and investigations.

Sources: https://www.justice.gov/usao-ndca/pr/cruise-admits-submitting-false-report-influence-federal-investigation-and-agrees-pay

Cruise robotaxi pedestrian dragging — What went wrong & fixes

- What went wrong: Safety case gaps; incident reporting deficiencies; inadequate post-market monitoring. **Remedy:** robust safety engineering, transparent reporting, regulator engagement.

Arup deepfake CFO video scam

Arup deepfake CFO video scam — One-line story

Employee wired roughly £20M after a realistic deepfake video call impersonated executives.

Source: https://www.theguardian.com/technology/article/2024/may/17/uk-engineering-arup-deepfake-scam-hong-kong-ai-video

Arup deepfake CFO video scam — What went wrong & fixes

- What went wrong: Weak out-of-band verification; missing dual-control on high-value transfers. **Remedy:** anti-fraud protocols, deepfake awareness, step-up verification.

Samsung temporary ChatGPT ban

Samsung temporary ChatGPT ban — One-line story

Sensitive code pasted into public AI tools triggered a temporary enterprise ban and policy reset.

Source: https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/

Samsung temporary ChatGPT ban — What went wrong & fixes

- What went wrong: Data loss prevention gaps; lack of secure, approved AI environments. **Remedy:** enterprise AI gateways, data classification, policy guardrails.

FTC Operation AI Comply crackdown

FTC Operation AI Comply crackdown — One-line story

U.S. FTC brought multiple cases against deceptive AI claims (so-called ‘AI washing‘).

Sources: https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes

FTC Operation AI Comply crackdown — What went wrong & fixes

- What went wrong: Unsupported claims; misleading marketing; fake capabilities. **Remedy:** substantiation standards, model evals, and clear disclosures.

SEC fines over AI washing

SEC fines over AI washing — One-line story

SEC penalized investment advisers for misleading AI claims, signaling securities-law risk for ‘AI washing‘.

Source: https://www.sec.gov/newsroom/press-releases/2024-36

SEC fines over AI washing — What went wrong & fixes

- What went wrong: False statements about AI capabilities. **Remedy:** compliance review of claims, documentation, and risk factors.

Controls: Observability, Security, Responsible AI

Observability pillar: logs & lineage

- Decision logs with input/output, version, features
- Data lineage from source $\rightarrow$ feature store $\rightarrow$ model
- Immutable audit trail; retention aligned to law/policy

Evaluation dashboards

- Pre-release: holdout + stress tests; post-release: shadow/canary
- KPIs: accuracy, latency, cost; KRIs: drift, hallucination, jailbreak rate
- Gates tied to thresholds; auto-rollbacks on breach

Fairness & bias monitoring

- Track TPR/FPR gaps; DIR/SPD; calibration across groups
- Confidence intervals; small-sample robustness
- Alerting + human review when thresholds trip

Drift detection

- Population Stability Index (PSI), KL divergence, KS tests
- $\mathrm{PSI} = \sum_i (p_i - q_i) \ln \frac{p_i}{q_i}$
- Trigger retraining & revalidation when drift persists

Model cards & change logs

- Use-case, risks, datasets, metrics, thresholds, owners
- Versioned changes with rationale and approvals
- Public summaries where required (EU AI Act transparency)

Security for LLMs (OWASP LLM Top-10)

- Prompt injection defenses; content filters; tool isolation
- Output verification; retrieval/domain whitelisting
- Red-team suites; jailbreak regression testing

Privacy engineering

- Data minimization; purpose limitation; k-anonymity/pseudonymization
- PIAs/DPIAs; restricted retention; secure enclaves
- Synthetic data & federated learning when feasible

Operational safeguards

- Kill-switches; rate limits; canary deployments
- Access control (least privilege); secrets isolation
- Incident response runbooks; simulator-based drills

Agentic AI: Why Autonomy Amplifies Risk

Why autonomy amplifies risk

- Tool calling couples text to action (payments, emails, code)
- Plans loop across multiple systems; error chains compound
- Attack surface expands: tool abuse, data exfiltration, SSRF via browsing

Deception & sleeper behavior

- Backdoors can persist through standard safety training
- Safety training may teach models to hide triggers
- Outcome: false sense of security if relying on refusal prompts

Guardrail brittleness

- Basic jailbreaks bypass safeguards across many models
- Agents inherit base model vulnerabilities and add new ones
- Continuous red teaming is mandatory, not optional

Containment patterns

- Hard permissions; bounded tools; allow/deny lists
- Output verification, typed tool schemas, dollar limits
- Staged rollouts, human approvals for sensitive actions

Agent incident response

- Comprehensive logs: plan, tools, outputs, approvals
- Immediate disable + rollback; notify stakeholders
- Root-cause + regression tests + policy/control updates

Industry Use Cases & Regulatory Overlays

Financial Services

Financial Services — Use Cases & Risks

- Credit underwriting, pre-qual, fraud triage, collections
- Risks: discrimination, explainability, adverse action, model risk

Financial Services — Regulatory Overlay & Controls

- Reg overlay: fair lending, adverse action reasons, NIST RMF mapping
- Controls: independent validation, bias audits, reject inference caution

Employment/HR

Employment/HR — Use Cases & Risks

- Screening, ranking, assessments, interview agents
- Risks: disparate impact, privacy, transparency/notice

Employment/HR — Regulatory Overlay & Controls

- Reg overlay: NYC LL 144; CA CPPA ADMT rights (2026)
- Controls: annual bias audit, notices, opt-out/appeal, data minimization

Healthcare

Healthcare — Use Cases & Risks

- Triage, diagnostics, coding, care navigation
- Risks: safety, bias by cohort, explainability, privacy (HIPAA)

Healthcare — Regulatory Overlay & Controls

- Reg overlay: quality/safety standards; FDA for certain devices
- Controls: clinical validation, human oversight, audit trails

Consumer Platforms

Consumer Platforms — Use Cases & Risks

- Moderation, recommender systems, content ranking
- Risks: misinformation, amplification, opacity

Consumer Platforms — Regulatory Overlay & Controls

- Reg overlay: EU DSA duties for VLOPs/VLOSEs
- Controls: risk assessments, transparency reports, choice architecture

Autonomy/Robotics

Autonomy/Robotics — Use Cases & Risks

- Perception, planning, control; warehouse automation; AV
- Risks: safety, real-world harm, liability

Autonomy/Robotics — Regulatory Overlay & Controls

- Reg overlay: safety cases, reporting, regulator audits
- Controls: simulation, scenario coverage, incident drills

Public Sector

Public Sector — Use Cases & Risks

- Benefits eligibility, risk scoring, chatbots, copilots
- Risks: due process, discrimination, transparency

Public Sector — Regulatory Overlay & Controls

- Reg overlay: procurement rules; M-24-10-era practices as references
- Controls: rights-impacting designations, appeals, logging

Key Regulations & Frameworks

EU AI Act

EU AI Act — Definition

Comprehensive EU regulation with risk-based tiers; entered into force 1 Aug 2024; prohibitions and AI literacy from 2 Feb 2025; GPAI & governance from 2 Aug 2025; fully applicable 2 Aug 2026; certain high-risk embedded systems extended to 2 Aug 2027.

EU AI Act — Intuition & Example

Intuition. Why it matters: extraterritorial reach and GPAI duties affect global providers; heavy documentation, transparency, post-market monitoring.

Example. Scope: high-risk use-cases, transparency duties, GPAI obligations, market surveillance. Timeline: see dedicated slides.

NYC Local Law 144 (AEDT)

NYC Local Law 144 (AEDT) — Definition

Requires annual independent bias audits, public posting of results, and candidate notices for automated employment decision tools; enforcement since 5 Jul 2023.

NYC Local Law 144 (AEDT) — Intuition & Example

Intuition. Why it matters: first mandatory bias-audit regime in the U.S., shaping hiring tools nationwide.

Example. Scope: hiring/promotion algorithms impacting NYC candidates/employees.

California CPPA ADMT Rules (CCPA)

California CPPA ADMT Rules (CCPA) — Definition

Adopted 24 Jul 2025; effective 1 Jan 2026. Mandates risk assessments, annual cybersecurity audits (for certain entities), ADMT access/opt-out rights, and enhanced disclosures.

California CPPA ADMT Rules (CCPA) — Intuition & Example

Intuition. Why it matters: de facto national standard for consumer ADMT rights; overlaps with vendor management and engineering evidence packs.

Example. Scope: California consumers' personal data; triggers for high-risk processing and ADMT usage.

Colorado AI Act (SB 24-205)

Colorado AI Act (SB 24-205) — Definition

High-risk AI duties for developers and deployers; effective date postponed to 30 Jun 2026 (from Feb 2026).

Colorado AI Act (SB 24-205) — Intuition & Example

Intuition. Why it matters: first broad state AI statute covering multiple domains (employment, credit, housing, education, healthcare).

Example. Scope: risk management program, impact assessments, incident reporting, notices/appeals.

EU Digital Services Act (DSA)

EU Digital Services Act (DSA) — Definition

Platform obligations for transparency, risk assessments, and mitigation for systemic risks (not AI-specific but impacts recommender systems).

EU Digital Services Act (DSA) — Intuition & Example

Intuition. Why it matters: content recommender governance and transparency duties affect ML systems at scale.

Example. Scope: online intermediaries; VLOPs/VLOSEs with heightened duties.

OECD AI Principles

OECD AI Principles — Definition

International principles for trustworthy AI (human-centered values, transparency, robustness, accountability).

OECD AI Principles — Intuition & Example

Intuition. Why it matters: sets global baseline adopted by many countries; useful for high-level governance alignment.

Example. Scope: Non-binding principles; reference for policy and corporate codes.

ISO/IEC 42001

ISO/IEC 42001 — Definition

AI Management System standard (AIMS) for organizational governance of AI (published 2023).

ISO/IEC 42001 — Intuition & Example

Intuition. Why it matters: certifiable governance scaffold that maps to laws and internal controls.

Example. Scope: policy, risk, lifecycle controls, supplier oversight, continual improvement.

NIST AI RMF 1.0 + GenAI Profile (NIST AI 600-1)

NIST AI RMF 1.0 + GenAI Profile (NIST AI 600-1) — Definition

Risk management framework and 2024 profile for generative AI; maps functions to concrete actions, measurements, and documentation.

NIST AI RMF 1.0 + GenAI Profile (NIST AI 600-1) — Intuition & Example

Intuition. Why it matters: U.S. consensus guidance trusted by regulators and auditors; highly actionable.

Example. Scope: Govern–Map–Measure–Manage functions across the model lifecycle.

OMB M-24-10 (Federal)

OMB M-24-10 (Federal) — Definition

Mar 28, 2024 memo—historic, binding requirements for federal AI governance and risk management across federal agencies (still a key reference for controls).

OMB M-24-10 (Federal) — Intuition & Example

Intuition. Why it matters: detailed operational guidance, inventories, impact/safety designations; widely reused by industry as a control library.

Example. Scope: U.S. federal government agencies and procurement.

U.S. EO 14179 (Jan 23, 2025)

U.S. EO 14179 (Jan 23, 2025) — Definition

Executive Order ‘Removing Barriers to American Leadership in AI’—revoked prior EO 14110 and refocused on innovation; directs an AI Action Plan.

U.S. EO 14179 (Jan 23, 2025) — Intuition & Example

Intuition. Why it matters: shifts federal posture but agency and state-level controls continue; does not preempt EU/state obligations.

Example. Scope: Federal policy direction, procurement emphasis, and innovation agenda.

EU AI Act — Application Timeline

- **1 Aug 2024**: Regulation enters into force
- **2 Feb 2025**: Prohibitions & AI literacy obligations apply
- **2 Aug 2025**: Governance rules & GPAI obligations apply
- **2 Aug 2026**: Broad applicability (most obligations)
- **2 Aug 2027**: Extended transition for certain high-risk systems embedded in regulated products

Source: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

U.S. State Timelines (selected)

- **NYC LL 144** (bias audits for AEDTs): enforcement since **5 Jul 2023**.
  https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page
- **California CPPA ADMT** regs adopted **24 Jul 2025**; effective **1 Jan 2026**.
  https://cppa.ca.gov/regulations/ccpa_updates.html
- **Colorado AI Act** (SB 24-205) effective date postponed to **30 Jun 2026**.
  https://leg.colorado.gov/bills/sb25b-004

U.S. Federal (context)

- **OMB M-24-10** (Mar 28, 2024): historic requirements for federal AI governance (still a key reference). https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-A pdf
- **EO 14179** (Jan 23, 2025): "Removing Barriers to American Leadership in AI" replaced prior EO 14110. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/

What This Means For You

Executives/Board

- Approve AI risk appetite; charter AI governance (AIMS/ISO 42001)
- Fund AI TRiSM; require dashboards (incidents, bias, drift, compliance)
- Name accountable owners and independent validation function

Product/Engineering

- Ship with gates: risk triage $\rightarrow$ threat/bias tests $\rightarrow$ HITL
- Log everything; implement kill-switches and rollback
- Secure tool-use; least privilege; rate limits; unit tests for prompts/tools

Data/ML Leaders

- Curate datasets; document lineage; quantitative fairness targets
- Pre/post-deployment evals; drift & bias monitoring; revalidation
- Model cards; change control; retrain triggers

Security/Privacy

- Threat models for LLMs; secrets isolation; RAG domain firewalls
- DLP; PIAs/DPIAs; vendor risk reviews; pentests/red-team exercises
- Incident response and breach playbooks for AI-specific failures

Legal/Compliance

- Map use-cases to applicable regs; maintain evidence packs
- Notices/consents/opt-outs; adverse action; audit rights in contracts
- Track multi-jurisdictional timelines; prepare for regulator inquiries

Managers/ICs

- Follow checklists; file model cards; run eval suites
- Escalate incidents; maintain documentation; continuous training
- Own remediation actions & deadlines after findings

Team Exercise

Exercise (Teams of 3–4)

Goal: Pick a use case → map applicable regulation(s) → select controls → make a recommendation.

Deliverable: 2-minute pitch + filled worksheet.

- Select a **use case** (candidate screener; credit pre-qual; refund agent; radiology assist; robotaxi routing)
- Identify **jurisdiction(s)** and **applicable rules**
- List **top risks**: safety, bias, privacy, security, fraud, accountability
- Choose **controls** (preventive/detective/corrective) and **evidence** to generate
- Define **metrics**: KPIs/KRIs & thresholds
- Give an **exec recommendation**: ship/hold; mitigations; timeline; owner

Worksheet Template (fill this in)

Use case	Jurisdiction(s) & key rules	Top risks	Controls (+ evidence

Summary & References

Key Takeaways

- Governance turns *principles* into *controls* and *evidence*
- Quantitative fairness $\neq$ one metric; choose by domain & base rates
- Observability is non-negotiable: logs, lineage, evals, drift, bias, incidents
- Agent autonomy multiplies risk; constrain tools, verify outputs, gate actions
- Multi-jurisdictional compliance is the new normal (EU + U.S. states)

Selected Sources (for further reading)

- EU AI Act timeline:
  https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
- NIST AI RMF + GenAI Profile:
  https://www.nist.gov/itl/ai-risk-management-framework
- ISO/IEC 42001: https://www.iso.org/standard/42001
- NYC Local Law 144:
  https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page
- CA CPPA ADMT Regs (2026 effective):
  https://cppa.ca.gov/regulations/ccpa_updates.html
- Colorado AI Act delay: https://leg.colorado.gov/bills/sb25b-004
- OWASP Top-10 for LLMs: https://genai.owasp.org/llm-top-10/
- Sleeper Agents (deceptive LLMs): https://arxiv.org/abs/2401.05566
- Equalized odds (Hardt et al., 2016): https://arxiv.org/abs/1610.02413
- Four-fifths rule (EEOC): https://www.law.cornell.edu/cfr/text/29/1607.4

Thank You

Questions?

Adnan Masood, PhD

amasood@amp207.hbs.edu — 626.513.1665