The Hard Edge of Trust: Governing Agentic AI

Risk, Regulation, and Real-World Controls

Guest Lecture by Adnan Masood, PhD.
Microsoft Regional Director — AI/ML Engineer — Author — Stanford Scholar
amasood@amp207.hbs.edu • +1 (626) 513-1665

University of South Florida — Graduate AI Seminar

October 9, 2025

Why this talk?

AI crossed from novelty to infrastructure.

- Agentic AI can *read, write, and act*: call tools, move money, write code, buy services, change systems.
- Governance is how we make that power **reliable, lawful, and fair** at scale.

Learning outcomes

1. Build a **taxonomy** of governance concepts and fairness metrics (with equations).
2. Recognize **why it matters** (1-line horror stories from real incidents).
3. Map **regulations** → **controls** you can implement now.
4. Practice with an **exercise**: use case → regulation → controls → recommendation.

This material is educational and not legal advice.

How we'll run this (interactive class format)

- **Term Triads**: Slide 1 = term only; Slide 2 = definition; Slide 3 = intuition + example.
- **Think–Pair–Share**: Short prompts; 60–90 seconds reflection; 2–3 minute table discussions.
- **Cold vs Warm Starts**: I'll ask for volunteers first, then call randomly.
- **Artifacts you keep**: AIA template, control catalog, fairness metrics cheat-sheet, role checklists.

Agenda

1. Taxonomy: governance & ethics
2. Quantitative fairness & bias
3. Why it matters (1-line horror)
4. Observability & controls
5. Regulations overview & timeline
6. Agentic AI risks
7. Impact by role (exec $\rightarrow$ dev)
8. Class exercise & recommendations

Governance

Governance: Definition

Governance is the system of *decision rights, accountability, and processes* that direct AI from idea to decommission, aligning with risk appetite, laws, and values. It answers *who decides, on what basis, with what evidence*.

Governance: Intuition & Example

Intuition. Org chart for AI decisions + the playbook to run them.

Example. An AI Steering Committee approves high-risk deployments; Model Risk signs off before launch; incident review board has kill-switch authority.

Compliance

Compliance: Definition

Compliance means *meeting binding obligations*: laws (e.g., EU AI Act), regulations, contracts, and internal policies/standards. Evidence is produced via documentation, testing, and audit trails.

Compliance: Intuition & Example

Intuition. Prove you did what you said and what was required.

Example. Keep a technical file, DPIA/AIA, data provenance, bias tests, human-oversight design, plus post-market monitoring logs.

Regulation

Regulation: Definition

Regulation is external rulemaking by governments/authorities that sets *minimum requirements, prohibitions, and enforcement*. Examples: GDPR Art. 22, EU AI Act risk tiers, China's deep synthesis rules, state bias-audit laws.

Regulation: Intuition & Example

Intuition. The floor, not the ceiling.

Example. Design controls to meet strictest applicable rule; reuse evidence across jurisdictions.

Audit

Audit: Definition

Audit is an independent, evidence-based assessment that *controls exist, are designed well, and operate effectively*. Can be internal, third-party, or regulatory (e.g., notified body).

Audit: Intuition & Example

Intuition. Trust, but verify—with artifacts.

Example. Auditor samples model versions, tests bias metrics, inspects logs, interviews owners, reproduces results.

Control

Control: Definition

A **control** is a specific *preventive, detective, or corrective* mechanism to mitigate a defined risk and meet an objective (policy). Controls are testable and owned.

Control: Intuition & Example

Intuition. Seatbelts, airbags, and crash reports for AI.

Example. Preventive: training data standards; Detective: drift monitors; Corrective: rollback + kill switch.

Ethical AI

Ethical AI: Definition

Ethical AI pursues *values* (fairness, beneficence, autonomy, justice) beyond bare compliance; often operationalized as *Responsible AI* programs with concrete practices.

Ethical AI: Intuition & Example

Intuition. "Should we," not only "can we."

Example. Decline a high-accuracy but privacy-invasive feature; add opt-out + consent and redesign data needs.

Fairness

Fairness: Definition

Fairness is the absence of unjustified, systematic disadvantage for protected groups. It's quantified by metrics like *statistical parity*, *equalized odds*, and *calibration within groups*.

Fairness: Intuition & Example

Intuition. Comparable error/benefit across groups for the task context.

Example. For hiring, ensure qualified candidates across demographics see similar true positive rates and low disparate impact.

Bias

Bias: Definition

Bias are systematic errors from data, labels, models, or deployment (sampling, measurement, historical, aggregation, evaluation). Distinguished from *intended policy* differences and *legally protected classes*.

Bias: Intuition & Example

Intuition. Garbage in, injustice out.

Example. Legacy credit data penalizes neighborhoods; model learns redlining proxies unless corrected.

Transparency

Transparency: Definition

Transparency reveals that AI is used and what it does: disclosures, documentation (model cards), and access to meaningful information for affected users or regulators.

Transparency: Intuition & Example

Intuition. No surprises; informed use.

Example. Chatbot explicitly states it is AI; deepfake content is labeled; publish summary of training data sources.

Explainability

Explainability: Definition

Explainability provides *human-understandable reasons* for outputs (e.g., global feature importances, local explanations). Fit-for-purpose: for developers, regulators, or end users.

Explainability: Intuition & Example

Intuition. Make the black box legible to the right audience.

Example. Provide top features and counterfactuals for a declined loan: "If income +\$5k, debt ratio ¡35%, decision flips."

Interpretability

Interpretability: Definition

Interpretability is model structure that is *intrinsically understandable* (e.g., sparse linear models, small trees) vs. post-hoc explanations for complex models.

Interpretability: Intuition & Example

Intuition. Simple when stakes allow; explain when complexity needed.

Example. Healthcare triage uses a sparse scorecard with clear thresholds.

Observability

Observability: Definition

Observability is end-to-end telemetry of AI systems: data lineage, model versions, evaluations, runtime metrics (accuracy, drift, bias, hallucination), and decision logs enabling *monitoring, diagnosis, and audit*.

Observability: Intuition & Example

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

Example. Dashboards track PSI drift, TPR/FPR by group, calibration error, toxicity, PII leaks, and tool-use traces for agents.

Human-in-the-loop

Human-in-the-loop: Definition

HITL embeds human judgment to approve, calibrate, or overturn AI decisions, with training, time, and authority to act; required in many high-risk settings.

Human-in-the-loop: Intuition & Example

Intuition. *Meaningful* oversight, not rubber stamps.

Example. A clinician must confirm an AI triage recommendation before action; overrides are logged and analyzed.

Algorithmic Impact Assessment (AIA)

Algorithmic Impact Assessment (AIA): Definition

AIA is a structured, pre-deployment risk assessment of an AI use case: context, stakeholders, harms/benefits, mitigations, tests, oversight, and post-market plan; maintained as a living artifact.

AIA: Intuition & Example

Intuition. An auditable design review for societal risk.

Example. Public benefits scoring AIA triggers bias testing, appeal routes, and strict logging requirements.

Quantitative fairness metrics (binary classification)

Confusion-matrix groups

	Actual 1	Actual 0
Pred 1	TP	FP
Pred 0	FN	TN

Statistical parity difference (SPD)

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

Disparate impact ratio (DIR, 80% rule)

$$\mathrm{DIR} = \frac{\Pr(\hat{Y} = 1 \mid A = a)}{\Pr(\hat{Y} = 1 \mid A = b)} \quad \text{(acceptable if } \approx 0.8\text{–}1.25\text{)}$$

Equal opportunity / TPR parity

Fairness trade-offs & selection

- **Trade-off theorem**: can't generally satisfy calibration, parity of error rates, and parity of base rates simultaneously.
- **Select metrics by context**: Lending (ECOA) often prioritizes disparate impact; hiring prioritizes equal opportunity; safety-critical prioritizes error asymmetry.
- **Set thresholds**: e.g., $|\mathrm{SPD}| \leq 0.05$, $\mathrm{DIR} \in [0.8, 1.25]$, $\Delta\mathrm{TPR} \leq 0.03$.
- **Mitigate**: reweighting, constraints, post-processing, feature review, label audit, policy changes.

Why governance matters: one-line horror stories (real incidents)

Welfare & public services

Algorithmic fraud scoring falsely flagged thousands of families; government resigned.

Hiring

A resume screener learned historical bias; women systematically down-ranked; tool scrapped.

Credit

A card limit model allegedly gave women far lower limits than comparable men; regulator inquiries ensued.

Policing

Predictive policing amplified over-policing in specific

From risk to controls: lifecycle view

Map risks to stages

- **Charter**: purpose, benefits/harms, risk appetite, lawful basis.
- **Data**: provenance, consent, quality, representativeness, PII.
- **Model**: design, eval plan, explainability, robustness, safety.
- **Deploy**: human oversight, safeguards, red-teaming, rollback.
- **Operate**: monitoring, drift, bias, incidents, retraining, sunset.

Control types

- **Preventive**: policies, standards, gating checklists, least-privilege.
- **Detective**: eval harnesses, canaries, SIEM hooks, fairness monitors.
- **Corrective**: kill switch, feature flags, incident playbooks.

Observability KPIs (examples)

Data: PSI < 0.2; missingness within spec; lineage 100% tracked; consent coverage $> 99\%$.
Model perf: AUC/accuracy by segment; ΔTPR/FPR across groups; ECE (calibration) < 0.02.
Safety: Adversarial robustness score; jailbreak detection rate; toxicity rate $< 0.1\%$.
Privacy: PII leakage rate $< 10^{-6}$/output; k-anonymity for logs; access audited.
Ops: MTTD/MTTR for drift; rollback < 15 min; on-call coverage; change management adherence.
Governance: AIA completion; sign-offs present; retraining cadence; incident postmortems completed.

Agentic AI

Agentic AI: Definition

Agentic AI performs multi-step plans with tool use and memory (browse, code, transact, control devices). Risks: *specification gaming*, *prompt injection*, *over-permissioned tools*, *data exfiltration*, *unsafe autonomy*.

Agentic AI: Intuition & Example

Intuition. "Software that writes software and executes it."

Example. A procurement agent drafts a contract, negotiates, and places an order—within spend and vendor constraints.

Why agents raise the stakes

- **Action surface**: Not just wrong text—*wrong actions*. Money moved; code deployed; systems altered.
- **Non-determinism**: Same prompt → different actions; requires guardrails and approvals.
- **Supply chain**: Models, prompts, tools, plugins, retrieval indices—each a risk node.
- **Security blend**: AppSec + MLOps + SecOps. Need *policy engines*, sandboxes, and ephemeral credentials.

Minimum controls for agentic AI (starter pack)

Policy & guardrails

- Allow-list tools; deny raw shell unless sandboxed.
- Action approval thresholds; two-person rule for high-risk.
- Rate limits; budget caps; scope-limited API keys.
- Content safety filters; PII scrub; watermarking outputs.

Observability & response

- Full *action logs* with inputs/outputs/artifacts linked.
- Real-time anomaly detection (policy violations).
- Evals for *goal drift*, hallucination, jailbreaking.
- Big red button: *pause agent* and revoke creds.

Key regulations & frameworks (selected)

Regime	Core ideas (non-exhaustive)
EU AI Act (2024–27)	Risk tiers (prohibited/high/limited/minimal); high-risk obligations: risk mgmt, data quality, technical file, human oversight, accuracy/robustness; GPAI/foundation-model duties; CE-like conformity; big fines.
GDPR Art. 22 (2018)	Limits on solely automated decisions with significant effects; rights to information and human review; DPIAs for high-risk processing.
China (2022–2023)	Algorithmic recommender rules; deep synthesis (labeling); generative AI measures (registration, content controls, security review).
U.S. (patchwork)	NIST AI RMF 1.0 (2023); Executive Order on AI (2023); NYC Local Law 144 (hiring bias audits); Colorado SB 205 (AI duties); sector enforcement (FTC/CFPB/FDA/EEOC).
UK (2023–)	Principles-first (safety, transparency, fairness, accountability, contestability) via sector regulators; AI Safety Institute for frontier risks.
Canada AIDA (proposed)	High-impact AI obligations (risk assessment, mitigation, incident reporting) with an AI/data commissioner.
OECD (2019), UNESCO (2021)	Global principles on trustworthy/rights-respecting AI; soft-law anchors adopted by many states.

Timeline (anchor milestones)

Year	Event
2018	GDPR in force (automated decision rights).
2019	OECD AI Principles adopted; Canada ADM Directive (AIA for gov).
2021	EU proposes AI Act; UNESCO global ethics recommendation.
2022	China algorithmic recommender rules in effect; U.S. AI Bill of Rights (blueprint).
2023	EU AI Act finalized; U.S. Executive Order on AI; NYC bias audit law effective; China generative AI measures.
2024–27	EU AI Act phased application (bans $\rightarrow$ GPAI $\rightarrow$ high-risk).
2026	South Korea AI Framework Act in force (high-impact focus).

Build for the strictest regime you face; reuse evidence across jurisdictions.

Map obligations to controls (EU AI Act → your backlog)

Obligation	Implementable control(s)
Risk management system	AIA template; harm catalog; sign-offs; risk register with owners/SLAs.
High-quality data	Datasheets; provenance; representativeness tests; label audits; bias-aware sampling.
Technical documentation	Model card; data card; evaluation reports; versioned pipelines; reproducibility scripts.
Human oversight	HITL workflow; appeal/override UI; training and competence evidence; RACI.
Accuracy/robustness	Eval harness; adversarial tests; stress tests; calibration; acceptance thresholds.
Post-market monitoring	Telemetry; drift & bias monitors; incident playbooks; retraining cadences.
Transparency	User notices; AI interaction badges; deepfake labels; accessible explanations.

Industry use cases where governance bites hardest

Sector	Typical AI Use	Governance pinch points
Healthcare	Diagnosis, triage, prior auth	Safety/efficacy, bias, informed consent, accountability, auditability.
Finance	Underwriting, AML, collections	Fair lending, explainability, model risk, adverse action notices.
Employment	Sourcing, screening, scheduling	Bias audits, candidate notice, disability accommodations, transparency.
Public sector	Benefits, fraud, policing	Fundamental rights, due process, appeal routes, proportionality.
Transport	Autonomy, dispatch	Functional safety, incident reporting, cybersecurity-by-design.
Platforms	Recommenders, moderation	DSA-like risk assessments, child safety, deepfake labeling, content harm.

Impact by role — what changes for *you*

Executives / Product

- Set risk appetite; empower an AI Risk Committee.
- Fund observability; require AIA sign-off pre-launch.
- Make metrics board-level (fairness, incidents, residual risk).

Engineers / Data Scientists

- Build eval harnesses; track metrics by segment.
- Design for explainability, rollback, and HITL.
- Red-team models; fix issues; document changes.

Managers

- Own control operation evidence; keep tecl file current.
- Staff on-call for AI incidents; run postmortems.

Security / Legal / Compliance

- Policy engine, sandboxing, least privilege, logging.
- DPIAs/AIA; vendor clauses; incident reporting.

Think–Pair–Share (1): Which fairness metric would you choose?

- **Use case**: University scholarship recommender; risk of excluding qualified low-income students.
- **Prompt**: Pick a metric (e.g., equal opportunity vs. parity) and defend it in 1 minute.

Think–Pair–Share (2): When should a human be in the loop?

- **Use case**: Oncology triage assistant suggesting care pathways.
- **Prompt**: Where is HITL mandatory, what authority, and what evidence proves "meaningful oversight"?

Class exercise (teams of 3–5)

1. **Pick a use case**: Hiring, credit, triage, policing, AV, content moderation, etc.
2. **Pick a regime**: EU AI Act, GDPR, NYC audit law, Colorado SB 205, China deep synthesis, NIST RMF.
3. **Find controls**: Preventive, detective, corrective—map to obligations.
4. **Recommend**: Ship / delay / cancel? What go-live gates and SLOs?

Exercise template (fill during breakout)

Field	Your entry
Use case	Domain, affected users, decisions made.
Harms/benefits	Top 3 potential harms; expected benefits.
Regime(s)	Cite the most constraining obligations.
Controls	For each obligation, list tests/telemetry/designs.
Ownership	RACI: who approves, who operates controls.
Go/no-go	Launch gates, SLOs, rollback plan, comms.

Summary — five takeaways

1. **Governance** = decision rights + process + evidence; **controls** make it real.
2. **Fairness is measurable** but contextual; pick metrics, set thresholds, test continuously.
3. **Observability** is non-negotiable: logs, evals, drift, incidents, and retraining.
4. **Regulations** converge on risk management, transparency, oversight, and documentation.
5. **Agents raise stakes**: add policy engines, allow-lists, approvals, and kill switches.

Artifacts you can reuse

- AIA template (risk register fields, sign-offs, tests).
- Control catalog (preventive/detective/corrective examples).
- Fairness metrics cheat-sheet (formulas & thresholds).
- Role checklists (Exec/Product/Manager/Engineer/Compliance).

Thank you

Questions? Discussion?

Adnan Masood, PhD.
amasood@amp207.hbs.edu Twitter/Blog/Books: search *Adnan Masood AI*

"Trust is a feature. Governance builds it."

Appendix: more bias sources

- **Sampling** (coverage, survivorship), **measurement** (sensor/label error), **historical** (societal patterns), **aggregation** (Simpson's paradox), **deployment** (population shift), **feedback loop** (performative effects).

Appendix: LLM/Agent eval ideas

- **Hallucination rate** (faithfulness to sources), **toxicity**, **prompt injection susceptibility**, **jailbreak success**, **PII leakage**, **tool-use accuracy**, **specification gaming tests**.

Appendix: Red-teaming menu

Threat	Test pattern
Prompt injection	Indirect injections via retrieved docs; HTML/CSV payloads; role confusion.
Data exfiltration	Secrets in prompts; credentials in env; RAG index leakage.
Unsafe actions	Overspend attempts; unsafe tool sequences; bypass approvals.
Content harm	Harassment, hate, self-harm, misinformation prompts.
Privacy	Quasi-identifier reconstruction; membership inference.

Appendix: DPIA vs AIA (quick contrast)

DPIA (privacy)	**AIA (algorithmic)**
Focus: personal data risk Law: GDPR/DP laws Artifacts: data flows, lawful basis Outcomes: mitigations, accept residual	Focus: decision/effect risk (incl. non-personal) AI-specific laws/policies (EU AI Act, state laws) Artifacts: metrics, evals, oversight, incidents Outcomes: go/no-go, launch gates, SLOs

Appendix: equations & thresholds (cheat-sheet)

- $\text{SPD} = \Pr(\hat{Y} = 1 | A = a) - \Pr(\hat{Y} = 1 | A = b)$. Target $|\text{SPD}| \leq 0.05$.
- $\text{DIR} = \Pr(\hat{Y} = 1 | A = a) / \Pr(\hat{Y} = 1 | A = b)$. Target $[0.8, 1.25]$.
- $\Delta\text{TPR}, \Delta\text{FPR} \leq 0.03$ where feasible; justify domain-specific deviations.
- $\text{ECE} = \sum_k \frac{|B_k|}{n} |\text{acc}(B_k) - \text{conf}(B_k)|$. Target < 0.02.