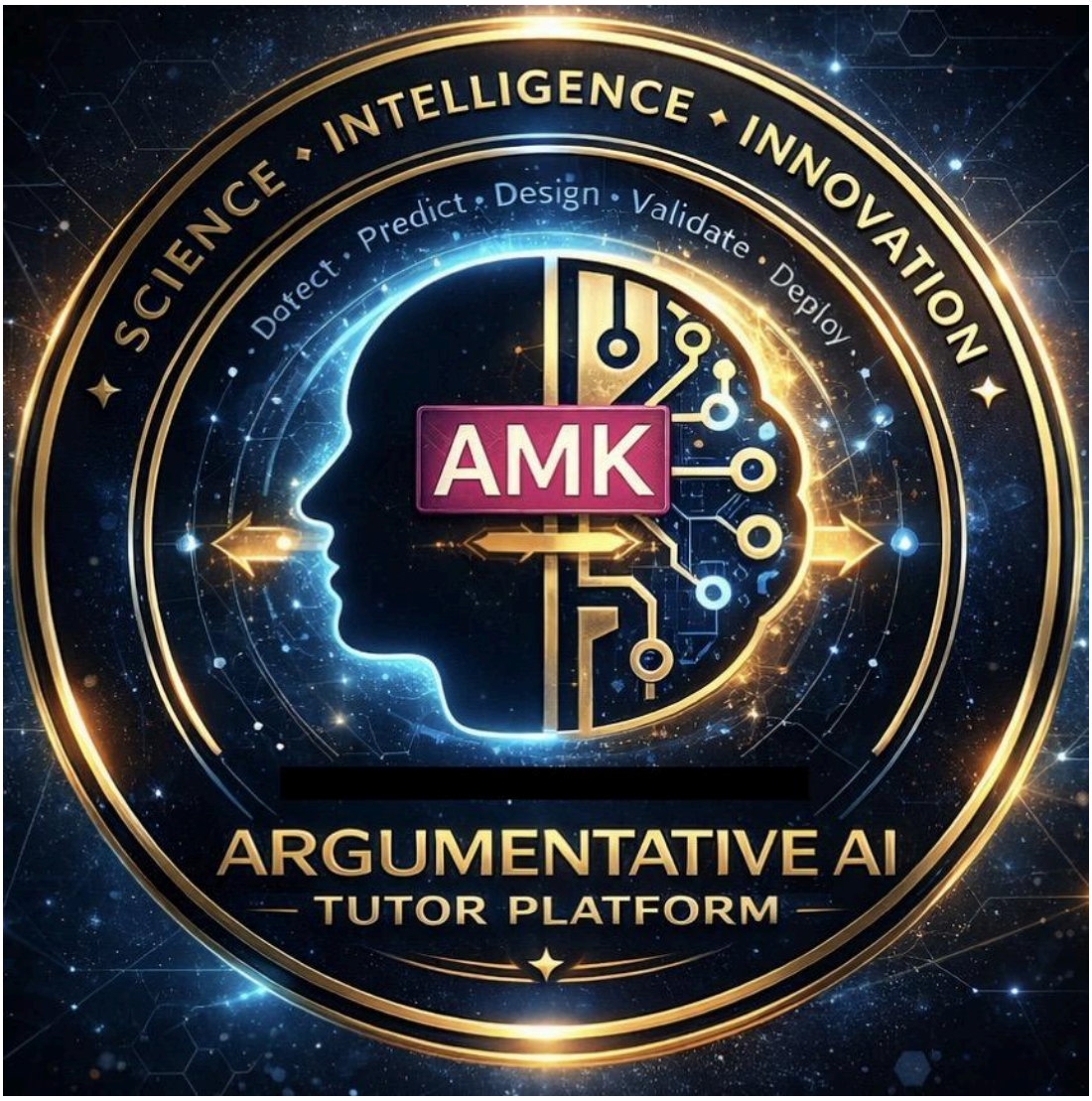

AI Certification Program Lecture Notes

Model-Level Risks Beyond Attacks, Breaches, Threats, and Vulnerabilities

With compliance, ethics, and control frameworks for trustworthy AI

Prepared for the AMK Argumentative AI Tutor Platform

> *Teaching point: an AI model can fail without being hacked. Many of the most serious risks come from design choices, data quality, misuse, over-automation, weak oversight, and poor governance.*

Learning outcomes

- Differentiate model-level risk from classic cybersecurity risk.
- Identify risks that emerge from data, objectives, autonomy, context, and human interaction.
- Apply NIST AI RMF, NIST GenAI Profile, OWASP guidance, MITRE ATLAS, OECD principles, UNESCO ethics guidance, ISO/IEC 42001, and the EU AI Act at a practical level.
- Map model risks to controls, governance processes, and evidence required for audit or certification.

Core idea. Traditional security focuses on attacks, breaches, threats, and vulnerabilities. AI governance must also address accuracy, robustness, explainability, fairness, privacy, accountability, and human oversight, as AI systems can cause harm even when no perimeter has been breached (NIST, 2023; NIST, 2024; OECD, 2024).

1. Why model-level risks matter

NIST describes AI risk as a combination of sociotechnical factors that can affect individuals, organizations, and society. That means risk is not limited to malware, unauthorized access, or system compromise. It also includes whether a model is fit for purpose, whether it is trustworthy in context, and whether human operators can understand and govern its outputs (NIST, 2023).

Examples of model-level risk beyond attacks:

- A language model fabricates legal or medical facts.
- A hiring model disadvantages protected groups.
- A recommendation model optimizes engagement but amplifies harmful content.
- A predictive model drifts because real-world conditions change.
- An autonomous system acts beyond the intended authority boundary.
- A team cannot explain why a high-impact decision was made.

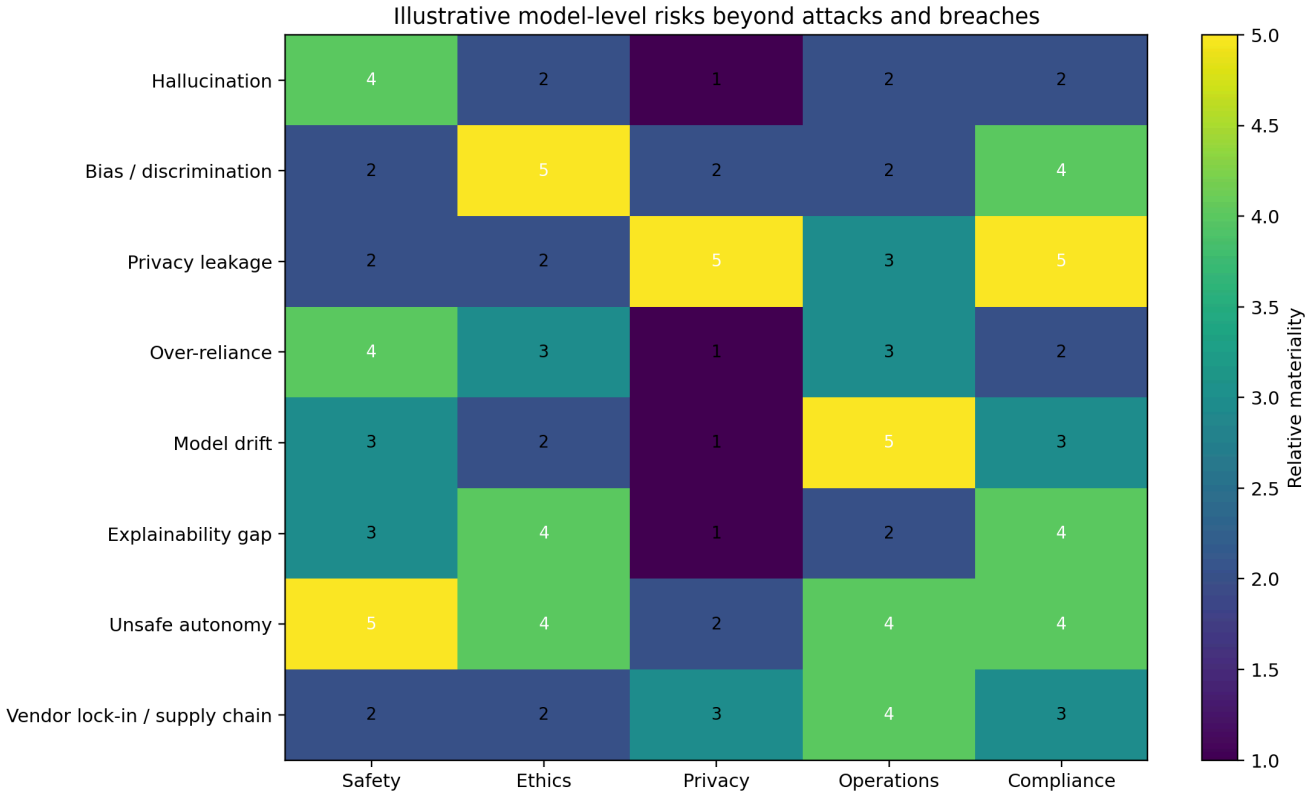

Figure 1. Instructional heatmap showing how model-level risks can materialize across ethics, privacy, operations, safety, and compliance domains.

2. Taxonomy of risks beyond attacks and breaches

2.1 Data risks
Poor labeling, skewed representation, stale training data, hidden proxies for sensitive attributes, and weak provenance can corrupt outcomes long before a cyber incident occurs.

2.2 Objective and reward risks
A model can optimize the wrong target. In practice, this appears as over-optimization of clicks, throughput, cost, or accuracy while neglecting safety, equity, or legal constraints.

2.3 Human-factor risks
Users may over-trust a model, misread confidence, or treat a draft as verified truth. NIST's GenAI Profile highlights over-reliance as a distinct governance problem (NIST, 2024).

2.4 Autonomy and agency risks
The more connected an AI system is to tools, plugins, workflows, or physical processes, the more important bounded authority becomes. OWASP flags excessive agency as a major risk for LLM-based systems (OWASP, 2025).

2.5 Lifecycle and operational risks
Drift, poor monitoring, inadequate fallback procedures, and weak incident response can turn a technically accurate model into an unsafe production system.

2.6 Third-party and supply-chain risks
Foundation models, APIs, datasets, plugins, and MLOps components introduce dependencies that can create hidden failure paths and compliance gaps (OWASP, 2025; ISO, 2023).

Risk-to-control mapping

Risk area	Typical failure mode	Primary control	Evidence
Data quality	Skew, label error, proxy bias	Dataset governance, data sheets, validation sampling	Lineage, validation reports
Hallucination	Confident false output	Retrieval grounding, verification workflow, human review	Prompt logs, QA records
Over-reliance	User accepts output without verification	UI warnings, confidence communication, workflow gates	User training, SOPs
Autonomy	Model triggers actions outside mandate	Least privilege, approval gates, kill switch	Access matrix, approvals
Drift	Performance declines after deployment	Monitoring, retraining triggers, rollback plan	Dashboards, incident tickets
Privacy leakage	Sensitive data exposed in prompts or outputs	Minimization, redaction, retention limits	Privacy impact assessment

Table 1. Example mapping from AI model risks to practical controls and auditable evidence.

3. Lifecycle view: where these risks emerge
The AI RMF organizes risk work into Govern, Map, Measure, and Manage functions. This lifecycle perspective is useful because many failures are introduced early but discovered late. A strong certification program teaches learners to ask not only "Can the model perform?" but also "How was it governed?" and "What evidence supports its deployment?" (NIST, 2023; ISO, 2023).

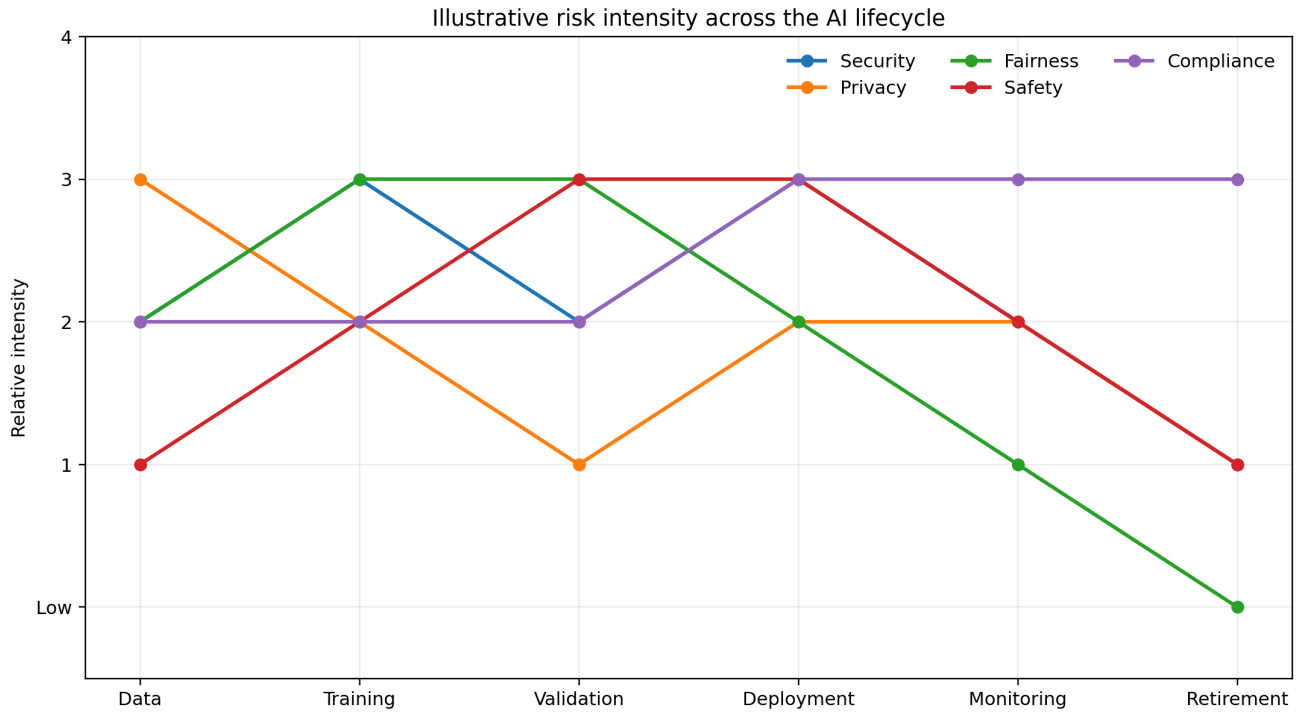

Figure 2. Instructional line chart showing that different risks peak at different points in the lifecycle.

Key lesson

- Most fairness and privacy failures begin upstream in data and problem framing.
- Most operational and autonomy failures appear during deployment and integration.
- Most compliance failures appear when controls are missing, undocumented, or not continuously reviewed.

4. Compliance, ethics, and governance frameworks

- **NIST AI RMF 1.0:** Provides a voluntary framework for trustworthy AI using the Govern, Map, Measure, and Manage functions.
- **NIST GenAI Profile:** Extends the AI RMF to generative AI and highlights risks such as confabulation, information integrity, privacy, harmful content, and over-reliance.
- **OECD AI Principles:** Promote innovative, trustworthy AI that respects human rights and democratic values; useful for leadership and policy alignment.
- **UNESCO Recommendation on the Ethics of AI:** Centers human rights, dignity, transparency, fairness, and human oversight.
- **ISO/IEC 42001:** Establishes an AI management system standard for lifecycle governance, risk management, and continual improvement.
- **EU AI Act:** Applies a risk-based regulatory model with stricter duties for high-risk AI use cases.
- **OWASP GenAI / LLM Top 10:** Offers implementation-focused guidance on common LLM and GenAI security failures such as prompt injection, insecure output handling, training data poisoning, model denial of service, excessive agency, and model theft.
- **MITRE ATLAS:** Provides AI-focused adversary knowledge for threat modeling across training, deployment, and inference.

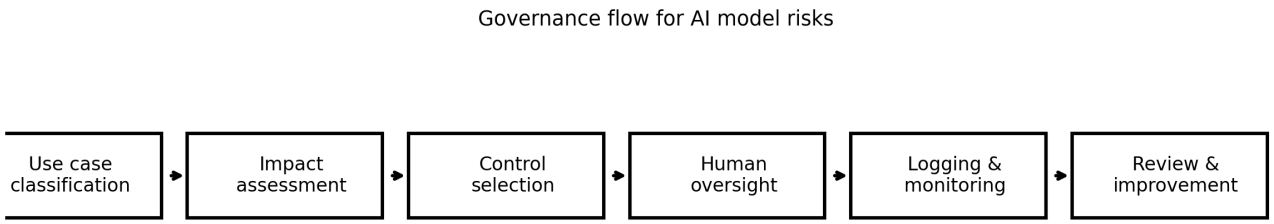

Figure 3. Governance workflow for selecting and operating controls over model-level risk.

5. Control families for certification

For teaching and assessment, it helps to group controls into families that learners can inspect, test, and document.

- Governance controls: policy, accountability, role ownership, model inventory, risk appetite, approval authority.
- Data controls: provenance, consent, minimization, quality checks, lineage, retention limits, bias testing.
- Model controls: evaluation metrics, benchmark selection, red-teaming, robustness tests, explainability artifacts, rollback paths.
- Human oversight controls: bounded autonomy, human approval gates, escalation triggers, override channels, training.
- Operational controls: monitoring, logging, incident management, service-level expectations, fallback procedures, decommissioning plans.
- Assurance controls: internal audit, independent review, documentation packs, change management, and post-deployment reviews.

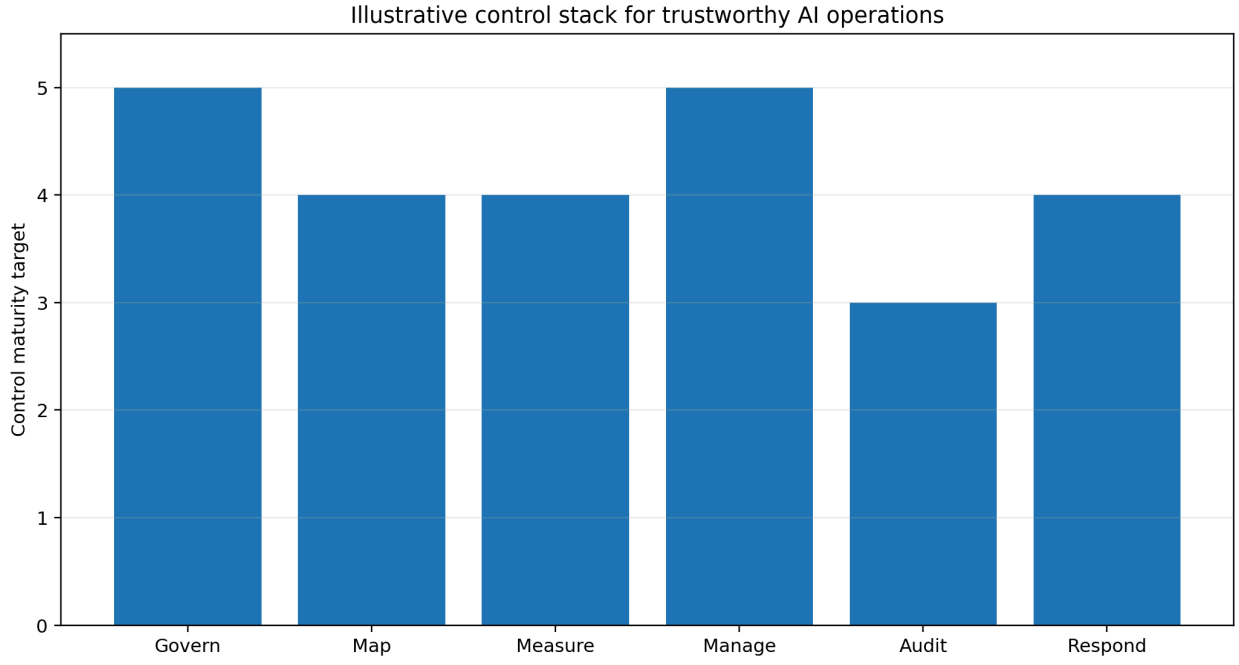

Figure 4. Instructional bar chart showing a balanced control stack for trustworthy AI operations.

6. Classroom and professional use cases

These notes fit an AI engineering, AI governance, cybersecurity, health informatics, or business analytics certification pathway. They can also support practical labs.

Use case	Main model-level risk	Control emphasis	Suggested lab
AI tutor	Hallucination and over-reliance	Verification workflow and teacher oversight	Compare raw vs verified answers
Hiring assistant	Bias and explainability gap	Impact assessment and review board	Adverse impact simulation
Clinical summarizer	Privacy leakage and unsafe advice	Minimum necessary access and human sign-off	Redaction and escalation exercise
Fraud model	Drift and false positives	Monitoring and threshold review	Drift dashboard lab

Table 2. Example classroom applications and labs for teaching model risk and control design.

7. Ten-question audit checklist

- What decision or action does the model influence?
- What harm can occur even if there is no cyberattack?
- Which groups could be unfairly affected?
- What personal or sensitive data enters prompts, training, or outputs?
- What human oversight exists before consequential action?
- How is performance measured in the real operating context?
- How are drift, misuse, and over-reliance detected?
- What happens when the model is uncertain or unavailable?
- Which laws, standards, and internal policies apply?
- What evidence would satisfy an auditor or regulator?

8. Instructor talking points

- Risk is broader than cybersecurity. AI governance must include technical, human, legal, and societal dimensions.
- Trustworthy AI requires controls before, during, and after deployment.
- Ethics is operational only when translated into roles, workflows, evidence, and monitoring.
- Compliance is not a substitute for safety, and safety is not a substitute for fairness; mature programs need all three.

References

- European Commission. (n.d.). Principles of the GDPR. European Commission.
- European Parliament. (2025). EU AI Act: first regulation on artificial intelligence. European Parliament.
- HHS. (2024). Summary of the HIPAA Security Rule. U.S. Department of Health & Human Services.
- ISO. (2023). ISO/IEC 42001:2023 Artificial intelligence — Management system. International Organization for Standardization.
- MITRE. (n.d.). ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. MITRE.
- NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology.
- NIST. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). National Institute of Standards and Technology.
- NIST. (n.d.). Privacy Framework. National Institute of Standards and Technology.
- OECD. (2024). OECD AI Principles. Organisation for Economic Co-operation and Development.

- OWASP. (2025). OWASP GenAI Security Project and Top 10 for LLM Applications. Open Worldwide Application Security Project.
- UNESCO. (2021/2024). Recommendation on the Ethics of Artificial Intelligence. United Nations Educational, Scientific and Cultural Organization.