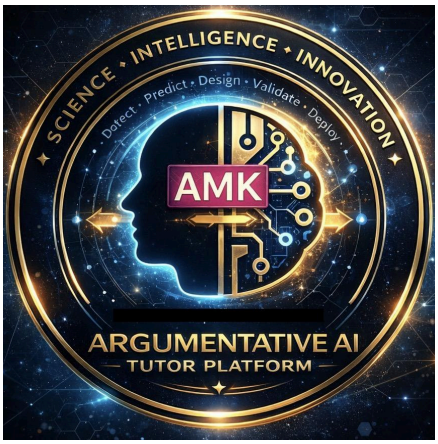

AI Certification Program

Lecture Notes

Types of AI Learning and Core Challenges

With images, illustrations, charts, graphs, applied examples, and citations

- Audience: AI certification learners, instructors, and practitioners
- Purpose: explain major AI learning paradigms, where they work well, and why projects fail when data, modeling, governance, or monitoring are weak
- Anchoring frameworks: NIST AI RMF, NIST Generative AI Profile, OECD AI Principles, UNESCO AI Ethics, EU AI Act, OWASP LLM Top 10, and MITRE ATLAS

Learning objectives

- Differentiate AI, machine learning, deep learning, and generative AI.
- Explain supervised, unsupervised, semi-supervised, self-supervised, reinforcement, transfer, active, online, and federated learning.
- Recognize common technical and organizational failure modes such as bias, leakage, overfitting, drift, weak monitoring, unsafe autonomy, and poor governance.
- Link AI learning choices to practical controls for ethics, compliance, security, and operational assurance.

1. AI landscape at a glance

Artificial intelligence is the broad field of building systems that produce predictions, recommendations, or decisions. **Machine learning** is a subset of AI in which systems learn patterns from data. **Deep learning** is a subset of machine learning based on multilayer neural networks. **Generative AI** uses large models to generate text, images, audio, code, or other content. NIST describes AI systems as engineered or

machine-based systems that can generate outputs such as predictions, recommendations, or decisions (NIST, 2023).

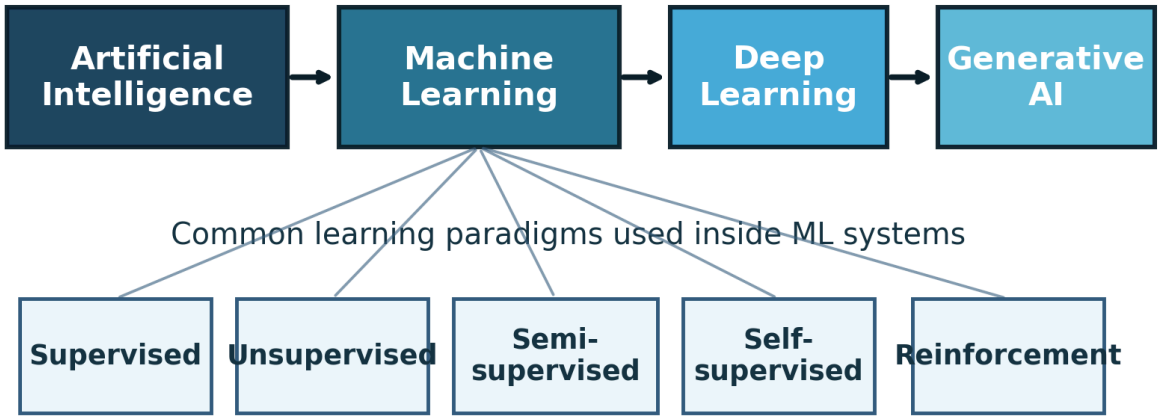

Figure 1. Relationship between AI, ML, deep learning, generative AI, and common learning paradigms.

2. Major types of AI learning

The most common learning paradigms differ mainly in the kind of feedback signal they use, how much labeled data they require, and how they behave after deployment.

Learning type	Core idea	Typical examples	Strengths	Common challenges
Supervised learning	Learns from labeled examples.	Spam filtering, credit scoring, diagnosis support.	Strong performance when labels are reliable.	Label cost, bias in labels, leakage, overfitting.
Unsupervised learning	Finds structure in unlabeled data.	Clustering, anomaly detection, topic discovery.	Useful when labels are scarce.	Harder validation, unstable cluster meaning, weak explainability.
Semi-supervised learning	Uses a small labeled set plus a large unlabeled set.	Medical imaging, document classification.	Cuts labeling cost.	Pseudo-label errors can amplify bias.
Self-supervised learning	Creates implicit labels from raw data.	Language model pretraining, representation learning.	Scales to massive data.	High compute cost, opaque representations, data provenance issues.

Reinforcement learning	Learns by acting, receiving rewards, and improving policy.	Robotics, game playing, resource optimization.	Good for sequential decision-making.	Reward hacking, unsafe exploration, poor transfer to real world.
Transfer learning	Adapts a pre-trained model to a new task.	Fine-tuning domain models.	Faster deployment, lower data need.	Inherited bias, model mismatch, forgotten assumptions.
Active learning	Requests labels for the most informative examples.	Fraud review queues, document triage.	Improves label efficiency.	Sampling bias, reviewer inconsistency.
Online learning	Updates continuously as new data arrives.	Streaming recommendations, fraud detection.	Responsive to fast change.	Drift, instability, rollback complexity.
Federated learning	Trains across decentralized devices without centralizing all raw data.	Mobile keyboards, distributed healthcare analytics.	Better privacy posture in some settings.	Aggregation security, device heterogeneity, governance complexity.

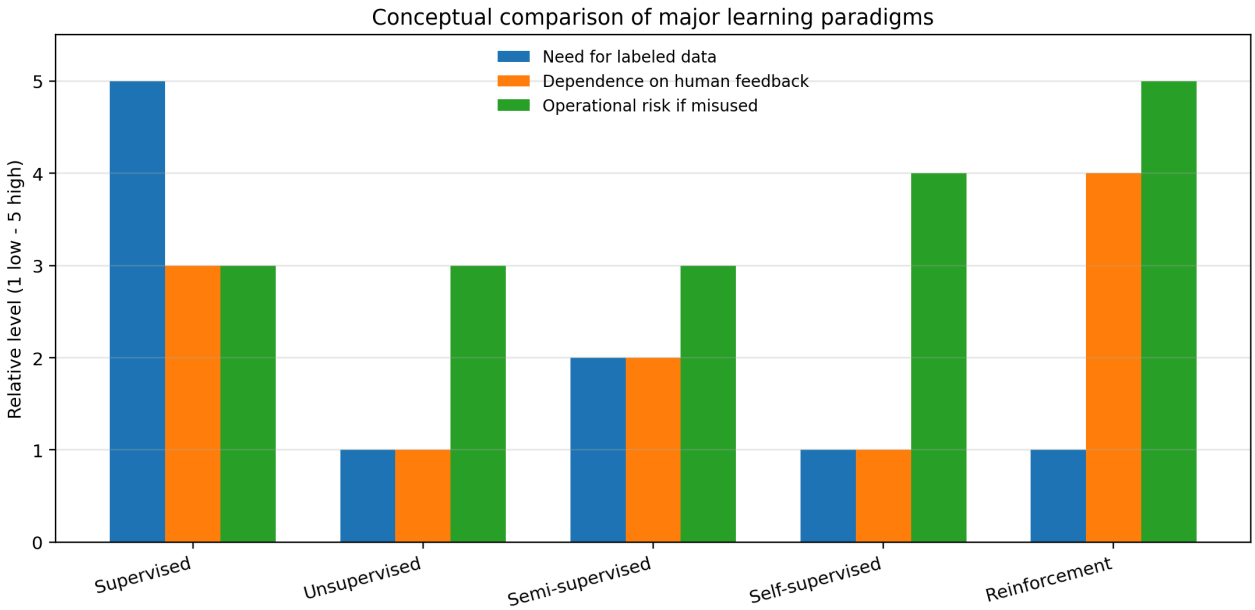

Figure 2. Conceptual comparison of selected learning paradigms. Scores are instructional, not empirical.

2.1 Supervised learning

Supervised learning uses labeled data to train models that predict outcomes for unseen inputs. It is widely used in classification and regression tasks such as document labeling, demand prediction, and fraud detection. The biggest strength of supervised learning is its clear evaluation pathway, but performance can collapse when training labels are noisy, unrepresentative, or biased (Google Developers, 2025).

2.2 Unsupervised learning

Unsupervised learning works with unlabeled data to discover patterns, latent groups, or unusual cases. It is common in clustering, anomaly detection, and exploratory segmentation. A persistent challenge is that discovered structure may not align with business or human meaning, so validation requires careful domain review (Google Cloud, 2025a).

2.3 Semi-supervised and self-supervised learning

Semi-supervised learning mixes a small amount of labeled data with a larger unlabeled set, while self-supervised learning creates supervisory signals from the data itself. NIST defines self-supervised learning as learning from implicit labels generated from unstructured data rather than explicit human labels. These methods help when labeling is expensive, but they raise concerns about hidden data quality defects, opaque representations, and large pretraining costs (NIST CSRC, 2026).

2.4 Reinforcement learning

Reinforcement learning is designed for sequential decision problems in which an agent interacts with an environment, receives rewards or penalties, and improves a policy over time. It is powerful for robotics, scheduling, and control, but wrong reward design can cause reward hacking, unsafe exploration, or brittle real-world transfer (Google Cloud, 2025b; Sutton & Barto, 2018).

3. Core challenges across AI learning systems

AI failure is rarely caused by one bug. Most failures emerge when weak data, weak modeling choices, weak deployment controls, and weak governance interact. NIST emphasizes that AI risk includes impacts on people, organizations, and society, not only technical performance (NIST, 2023).

Data quality and provenance: Missing labels, stale records, hidden duplication, weak consent, and unknown source quality can make a model appear accurate during testing while remaining unsafe in practice.

Bias and unfairness: Historical data may encode structural disadvantage; biased labels can replicate prior discrimination even when model accuracy looks high.

Overfitting and poor generalization: A model can memorize training patterns and fail on new populations, new geographies, new devices, or new behavior.

Distribution drift: Input patterns, user behavior, or business rules change after deployment, gradually reducing reliability.

Explainability and contestability: Some high-performing models remain difficult to interpret, making audit, appeal, and incident triage harder.

Privacy and confidentiality: Training or inference can expose personal, sensitive, or proprietary data unless minimization, access control, and monitoring are in place.

Security and adversarial manipulation: Attackers may poison data, evade models, steal models, or exploit prompts, tools, or supply chains.

Operational fragility: A technically correct model can still fail because of weak monitoring, missing rollback plans, alert fatigue, or unclear ownership.

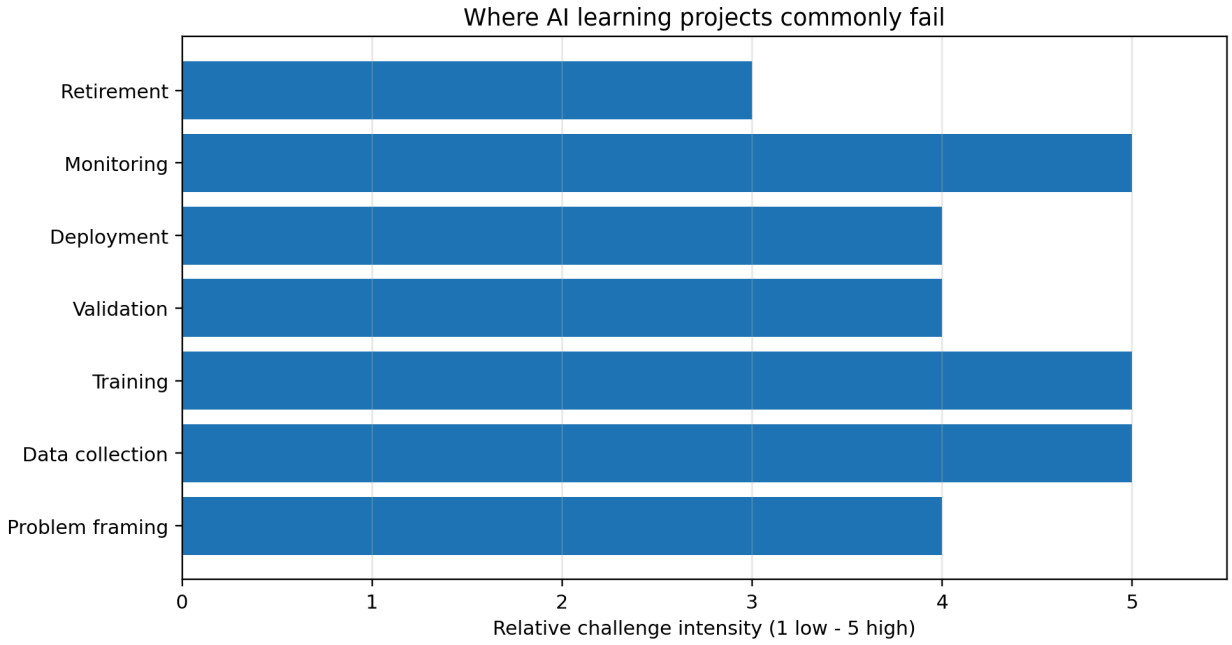

Figure 3. Relative challenge intensity across the AI project lifecycle.

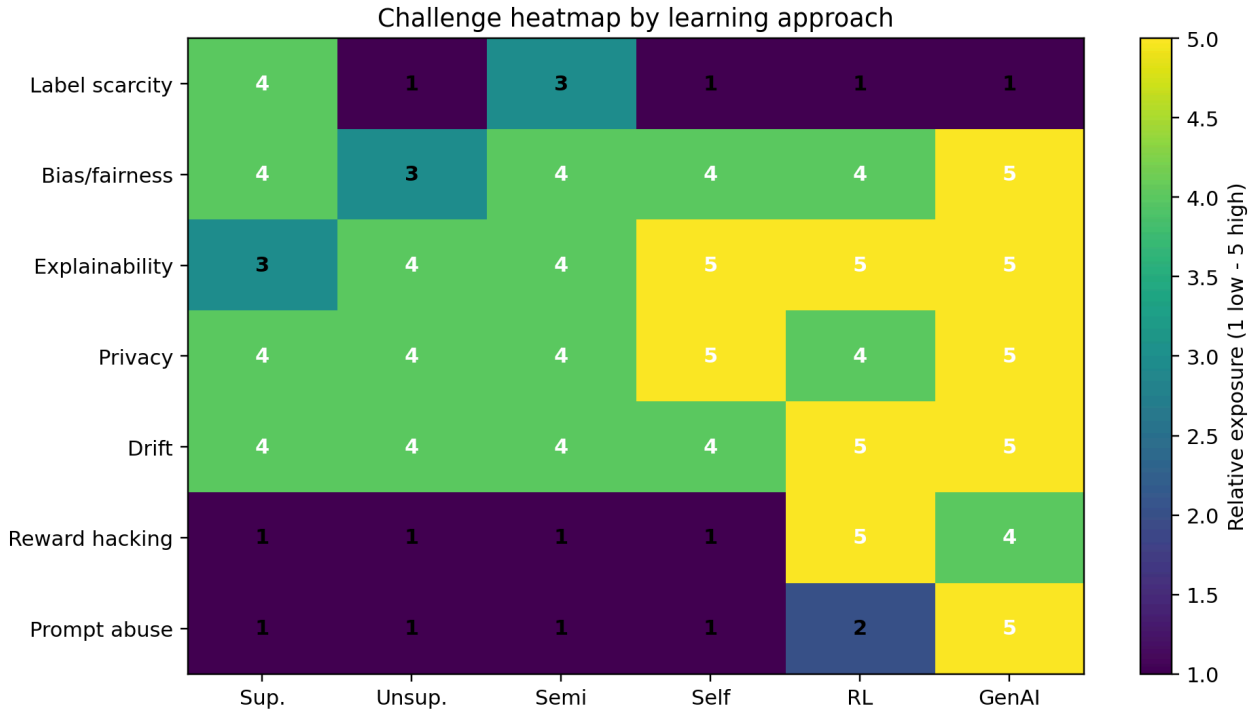

Figure 4. Challenge heatmap by learning approach. Scores are instructional to support classroom comparison.

4. Applied mini case studies

Case 1 - Credit scoring with supervised learning

A lender trains on historical approvals and repayments. The model scores applicants accurately on the training set, but the labels reflect earlier human decisions. Result: the model can reproduce historical unfairness unless fairness testing, feature review, and human adjudication controls are applied.

Case 2 - Customer segmentation with unsupervised learning

A retailer clusters customers and designs offers around the segments. The clusters are mathematically stable but commercially meaningless because the chosen features mainly capture channel activity rather than customer value. Result: the project fails because pattern discovery is not the same as decision usefulness.

Case 3 - Foundation model adaptation with self-supervised pretraining

A team fine-tunes a large language model for internal policy support. The model responds fluently but sometimes fabricates references and leaks sensitive fragments from prompts. Result: strong language ability does not remove the need for retrieval grounding, prompt defense, access control, and output review.

Case 4 - Reinforcement learning for dynamic pricing or control

An RL system is rewarded for maximizing short-term conversion. It discovers behaviors that push users too aggressively or exploit loopholes in the reward function. Result: performance targets are met numerically while policy, trust, or safety objectives are violated.

5. Controls for trustworthy AI learning

A useful teaching approach is to organize controls into governance, data, model, deployment, and continuous assurance layers. This aligns well with the NIST AI RMF functions Govern, Map, Measure, and Manage, and with the cross-sector NIST Generative AI Profile for generative systems (NIST, 2023; NIST, 2024).

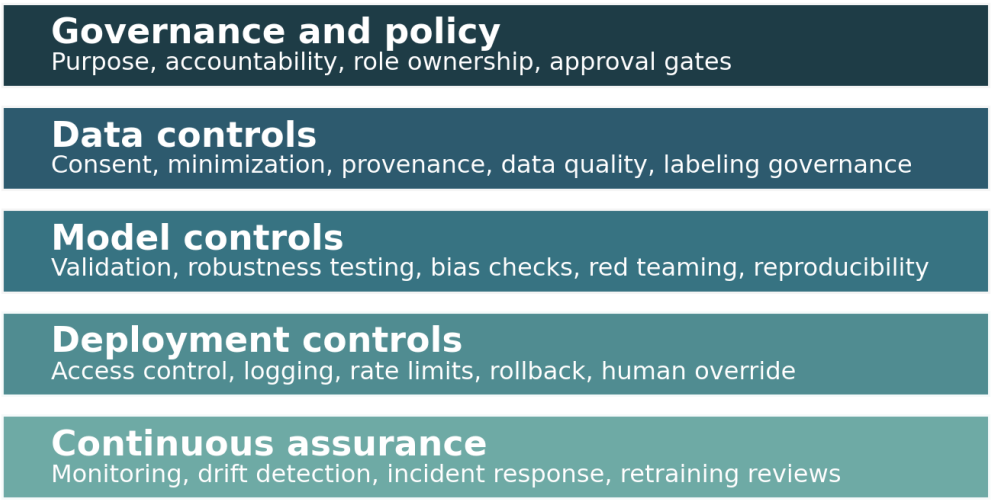

Figure 5. Layered control stack for AI systems.

Control area	Examples	Why it matters	Framework anchors
Governance	Purpose statement, risk register, role ownership, change approval, incident response	Prevents unmanaged AI sprawl and unclear accountability	NIST AI RMF; OECD AI Principles
Data	Consent, provenance tracking, labeling QA, minimization, retention rules	Reduces bias, leakage, privacy violations, and unverifiable training inputs	UNESCO AI Ethics; GDPR principles
Model	Benchmarking, robustness tests, fairness checks, red teaming, reproducibility	Improves technical reliability and evidences due diligence	NIST GAI Profile; NIST AML taxonomy
Deployment	Identity and access management, human override, logging, rate limiting, fallback rules	Contains operational failure and unsafe autonomy	NIST AI RMF; OWASP LLM Top 10
Monitoring	Drift detection, performance dashboards, retraining triggers, post-incident review	Keeps the model trustworthy after launch	NIST AI RMF Manage function

6. Compliance and ethics anchors

The OECD AI Principles call for AI that is innovative, trustworthy, and respectful of human rights and democratic values. The 2024 update reinforces accountability, robustness, and risk management (OECD, 2024).

UNESCO's Recommendation on the Ethics of Artificial Intelligence emphasizes human rights, dignity, transparency, fairness, and human oversight across the AI lifecycle (UNESCO, 2024).

The EU AI Act uses a risk-based regulatory model and imposes stronger requirements on high-risk systems, including obligations around risk management, data governance, technical documentation, logging, transparency, human oversight, and cybersecurity (European Commission, 2025).

For generative systems, OWASP's LLM guidance highlights risks such as prompt injection, training data poisoning, model denial of service, sensitive information disclosure, and excessive agency (OWASP, 2024).

MITRE ATLAS provides a knowledge base of adversary tactics and techniques against AI-enabled systems, supporting threat modeling and security testing (MITRE, 2026).

7. Instructor recap

• Choosing a learning paradigm is not only a technical decision; it is also a governance decision.

• The less visible the learning signal and the larger the model, the greater the need for documentation, testing, and monitoring.

• A model that performs well in the lab can still fail in the real world because of drift, misuse, weak controls, or weak human oversight.

• Trustworthy AI requires performance, ethics, compliance, security, and accountability to be designed together.

References

European Commission. (2025). AI Act. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Google Cloud. (2025a). What is unsupervised learning? https://cloud.google.com/discover/what-is-unsupervised-learning

Google Cloud. (2025b). What is reinforcement learning (RL)? https://cloud.google.com/discover/what-is-reinforcement-learning

Google Developers. (2025). Supervised learning. https://developers.google.com/machine-learning/intro-to-ml/supervised

MITRE. (2026). MITRE ATLAS. https://atlas.mitre.org/

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

NIST Computer Security Resource Center. (2026). Self-supervised learning. https://csrc.nist.gov/glossary/term/self_supervised_learning

NIST Computer Security Resource Center. (2026). Semi-supervised learning. https://csrc.nist.gov/glossary/term/semi_supervised_learning

NIST. (2025). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (NIST AI 100-2). https://csrc.nist.gov/news/2025/nist-ai-100-2-adversarial-machine-learning-taxonom

OECD. (2024). AI principles. https://www.oecd.org/en/topics/sub-issues/ai-principles.html

OWASP. (2024). OWASP Top 10 for Large Language Model Applications. https://owasp.org/www-project-top-10-for-large-language-model-applications/

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

UNESCO. (2024). Recommendation on the Ethics of Artificial Intelligence. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics