AI Foundations

Certified Responsible AI Governance and Ethics Professional

KNOWLEDGE CHEATSHEET

FROM DATA TO DECISIONS

Data Prepare → **Learning** Train → **Validation** Evaluate → **Decision** Oversee

01 AI foundations and applications

Learning and analytics

- **Supervised learning** uses labeled examples; **unsupervised learning** discovers structure without target labels. **Semi-supervised learning** combines labeled and unlabeled data.
- **Reinforcement learning** learns a policy through actions, environmental feedback and rewards. Poorly specified rewards can encourage behavior that meets the score while defeating the real objective.
- **Descriptive analytics** asks what happened; **diagnostic** asks why; **predictive** estimates what may happen; **prescriptive** recommends actions. Historical association alone does not establish causation.
- **Medical imaging and diagnostics** analyzes images for clinically relevant patterns. Treat model outputs as decision support within a validated clinical workflow, with appropriate specialist oversight.

Architectures and AI categories

- **Feedforward networks** map inputs to outputs without recurrent state. **CNNs** exploit spatial structure, making them useful for image tasks.
- **RNNs** maintain state across sequences. **Transformers** use attention to relate sequence elements and support parallel training; autoregressive generation still produces tokens sequentially.
- **Autoencoders** learn compressed representations through reconstruction. **GANs** train a generator against a discriminator to produce synthetic examples.
- **Reactive machines** respond to the current state without learning from past interactions. **Limited-memory AI** uses historical information; theory-of-mind and self-aware categories are conceptual, not ordinary deployed capabilities.

Language preprocessing

- **Tokenization** splits text into tokens such as words or subwords. **Stemming** heuristically trims word forms; **lemmatization** maps them to dictionary base forms.
- **Sentiment analysis** estimates expressed attitude. Preprocessing, language coverage and context affect reliability; sarcasm and domain-specific meanings can defeat simple keyword rules.

Ethics & Governance

Certified Responsible AI Governance and Ethics Professional

KNOWLEDGE CHEATSHEET

GOVERNANCE IN PRACTICE

Values Principles → **Policy** Expectations → **Roles** Authority → **Controls** Evidence

02 Ethics and governance foundations

People, fairness and accountability

- **Ethical concerns** involve values such as dignity, autonomy and equal opportunity even when no law has been broken. **Societal concerns** include effects on employment, institutions and communities.
- **Human-centric values** preserve agency, participation and meaningful choices. Give affected people notice, accessible explanations, review routes and ways to contest consequential outcomes.
- **Fairness** requires assessing who benefits and who bears errors or burdens. Compare relevant groups and contexts; high overall accuracy can hide unequal false-positive or false-negative rates.
- **Accountability** assigns identifiable owners for decisions and remediation. Human oversight needs training, time, authority and the practical ability to override or stop unsafe automation.
- **Safety** addresses accidental harm; **security** addresses hostile compromise. **Sustainability** includes environmental resource use and the ability to maintain dependable operation over time.

Bias sources and feedback loops

- **Data bias** can arise from sampling, missing populations, labels or measurement. **Algorithmic bias** can arise from objectives, feature choices, optimization and decision thresholds.
- **Societal bias** reflects existing structural inequalities. **Confirmation bias** influences human selection and interpretation of evidence. Multiple sources can coexist in one system.
- **Balanced datasets** do not guarantee fair outcomes: labels, proxy variables, subgroup interactions and deployment conditions still matter. Evaluate outcomes as well as data composition.
- **Feedback loops** occur when model decisions shape later training data, reinforcing earlier preferences. Review opportunities denied, not just the success of people selected.

Framework purposes

- **OECD AI Principles** guide trustworthy AI through human rights, fairness, transparency, robustness and accountability. They support governance alignment without replacing applicable laws.
- **UNESCO AI ethics guidance** emphasizes human rights, inclusion, cultural diversity, participation and environmental well-being across the AI lifecycle.

Ethics & Governance / Strategy & Vendors

Certified Responsible AI Governance and Ethics Professional

KNOWLEDGE CHEATSHEET

GOVERNANCE IN PRACTICE

Values Principles → **Policy** Expectations → **Roles** Authority → **Controls** Evidence

- **IEEE 7000-series standards** connect ethical concerns to engineering processes. **DoD AI Ethical Principles** emphasize responsible, equitable, traceable, reliable and governable use in defense.
- **Framework selection** considers organizational goals, system impact, legal obligations, flexibility and scalability. Combine complementary guidance with measurable controls suited to the use case.

Governance operating model

- **Values and principles** express intent; policies define expectations; **institutional structures** assign oversight; operational controls make requirements repeatable and auditable.
- **Governance boards and ethics panels** review significant risks and approve exceptions. Document decision rights, stop authority, escalation thresholds and the owner of residual-risk acceptance.
- **Compliance protocols** translate obligations into lifecycle gates, evidence requirements and repeatable checks. Connect approvals, validations, changes and reviews through a common system identifier.
- **Responsibility matrices** distinguish business ownership, development, independent validation and live operation. **Data stewards** govern data quality and lineage; operators supervise outputs and escalate anomalies.
- **Stakeholder mapping** identifies affected groups, influence, interests and exposure to harm. Include people with limited decision power who may experience substantial consequences.

03 AI strategy and third-party oversight

Readiness, prioritization and scaling

- **Readiness assessment** examines data, skills, infrastructure, processes and governance. Manual pipelines, isolated pilots and inconsistent controls indicate early maturity; maturity labels vary by model.
- **Use-case prioritization** balances business value, feasibility, strategic alignment and risk. High expected impact does not make a project a quick win when data or infrastructure is unavailable.
- **Roadmaps** sequence initiatives, dependencies, milestones, owners and success measures. **Phased value realization** builds experience through manageable deployments before higher-risk expansion.
- **Pilot evaluation** considers predictive quality, business value, usability and workflow adoption. Set success and stop criteria in advance, then decide whether to pivot, persevere or stop.
- **Hub-and-spoke scaling** combines central standards, platforms and oversight with departmental implementation. Use phased expansion, interoperability and feedback to grow without fragmented controls.

Strategy & Vendors / Risk & Security

Certified Responsible AI Governance and Ethics Professional

KNOWLEDGE CHEATSHEET

RESPONSIBLE ACQUISITION

Need Define value → **Vendor** Assess risk → **Contract** Set controls → **Service** Monitor

Procurement and vendor risk

- **Procurement begins with problem clarity**: define the need, expected value, affected users and acceptable risk before selecting a product. Translate these into technical and assurance requirements.
- **Cloud providers** operate infrastructure; **model providers** supply models; **AI API providers** expose hosted capabilities; **data vendors** supply datasets. Assign responsibilities to the actual service boundary.
- **Due diligence** examines governance, data provenance, model integrity, security, privacy, validation evidence, operational resilience and vendor track record before onboarding.
- **Risk tiers** reflect criticality, sensitive data, dependency and potential harm. Mission-critical vendors warrant stronger evidence, monitoring and continuity planning; tier numbers are organization-specific.
- **Third-party risk management** continues through integration, ongoing monitoring, change review and exit. Outsourcing a capability does not remove the organization's accountability for its use.

Contracts that support oversight

- **Transparency requirements** specify model documentation, intended use, limitations, change notices and access to decision explanations. **Audit rights** enable verification of contractual and regulatory commitments.
- **Human oversight and incident clauses** define intervention rights, notification deadlines, cooperation, escalation and remedies. Cover subprocessors, data ownership, permitted use and return or deletion at exit.
- **SLAs** establish service commitments and consequences for failure. **SLOs** are measurable objectives such as availability or latency; define response, recovery and resolution terms separately.
- **Cross-border processing** requires mapping storage, access, support and subprocessor locations. Check transfer mechanisms, residency constraints and retention obligations for each applicable jurisdiction.

04 Risk, privacy and security controls

Risk assessment and prioritization

- **Risk identification** finds hazards and failure scenarios. **Analysis** evaluates likelihood, severity, affected people and cascading consequences; **prioritization** compares exposure with defined tolerance.
- **Vulnerability** is a weakness that can be exploited; **likelihood** concerns occurrence; **impact** concerns consequences. Record inherent risk, controls, residual risk, owners and review triggers.

Risk & Security

Certified Responsible AI Governance and Ethics Professional

KNOWLEDGE CHEATSHEET

RISK MANAGEMENT CYCLE

Govern Set direction → **Map** Understand → **Measure** Evaluate → **Manage** Respond

- **Bow-tie analysis** links causes, a central event and consequences with preventive and mitigative barriers. **Fault trees** trace causes; **event trees** explore outcomes; **Monte Carlo simulation** samples modeled uncertainty.
- **NIST AI RMF**: **Govern** establishes accountability and policy; **Map** understands context and impacts; **Measure** evaluates risk; **Manage** prioritizes responses and allocates resources. Govern spans the other functions.
- **Impact assessments** examine affected rights and outcomes. **Trust scoring** can summarize repeatable criteria, but retain component scores and evidence so an average does not conceal a serious weakness.

Privacy throughout the data lifecycle

- **Privacy by design** limits collection, purpose, access and retention. Map ingestion, training, inference, logs and backups; protect sensitive outputs as well as source records.
- **GDPR erasure rights** apply under specified conditions and include exceptions. Plan retention and storage so valid deletion requests can be honored rather than blocked by immutable architecture.
- **HIPAA** protects health information within its covered-entity and business-associate scope. Identify applicable privacy and security duties and required agreements; health-related data is not automatically within HIPAA.
- **Pseudonymization** replaces identifiers while permitting re-linking with additional information. **Anonymization** aims to prevent identification; masking alone does not establish anonymity.
- **Federated learning** shares model updates while keeping raw training data local. **Edge processing** runs computation near the source. Neither automatically eliminates leakage, security or legal risks.
- **Encrypted computation**, such as homomorphic encryption, enables supported calculations without exposing plaintext to the computing party. Evaluate cost and supported operations; synthetic data still needs privacy validation.

Secure architecture and access

- **Defense in depth** protects data, models, applications and infrastructure through independent layers. **Security by design** and a **secure development lifecycle** embed controls from ingestion through retirement.
- **Zero trust** avoids implicit trust based on network location and continually evaluates access context. **Least privilege** limits each identity to needed actions and resources.
- **Micro-segmentation** isolates workloads to limit lateral movement. Combine identity-aware policy, restricted service communication and monitored administrative access.
- **Repository security** combines role-based access, protected changes, code review and audit logs for datasets, models and code. Automate integrity and security checks in release pipelines.

Risk & Security / Testing & Deployment

Certified Responsible AI Governance and Ethics Professional

KNOWLEDGE CHEATSHEET

RISK MANAGEMENT CYCLE

Govern Set direction → **Map** Understand → **Measure** Evaluate → **Manage** Respond

Inference APIs and behavioral guardrails

- **API protection** requires authenticated callers, authorization, transport encryption and secret management. Validate token signatures, issuer, audience and expiry; OAuth 2.0 governs delegated authorization and JWT is a token format.
- **Rate limits, quotas and DDoS protection** reduce resource exhaustion and abuse. Monitor unusual requests and sensitive-output exposure; throttling alone does not make incorrect predictions safe.
- **Guardrails** constrain approved use: contextual controls preserve task boundaries, policy controls enforce rules, and safety or ethical controls address harmful behavior. Test interactions among controls.
- **Context-aware filtering** evaluates conversation history and intended use. Combine pattern rules, semantic classification and human review to manage nuanced harm and false positives.

AI threats and control frameworks

- **Training-data poisoning** corrupts examples or labels to alter learned behavior. **Model tampering** changes artifacts or parameters; **adversarial evasion** crafts inference inputs to mislead a model.
- **Model extraction** attempts to reproduce a model through access to its outputs. **Privacy attacks** may infer training membership or sensitive information; control access, disclosure and suspicious query patterns.
- **Supply-chain risk** includes compromised dependencies, datasets and externally supplied models. Validate provenance and integrity independently; opaque development practices increase assurance gaps.
- **STRIDE** covers Spoofing, Tampering, Repudiation, Information disclosure, Denial of service and Elevation of privilege. Missing attributable logs create repudiation risk.
- **CSA AI Controls Matrix (AICM)** organizes AI security and governance controls. **OWASP AI and LLM security guidance** highlights application threats and mitigations; neither a checklist nor a vendor claim proves compliance.

05 Testing, explanation and deployment

Risk-based test planning

- **Define testing scope** across data, models, pipelines, interfaces and downstream decisions. Prioritize high-impact decisions, safety and security, then analyze failure, misuse and abuse scenarios.
- **Functional tests** verify required behavior; fairness tests compare relevant groups; security tests probe hostile behavior; **UAT** checks real workflows and user acceptance. Coordinate these streams before release.
- **Training-data verification** checks provenance, permissions, representativeness, labels, missing values and leakage. Separate training, tuning and independent evaluation data.

Testing & Deployment

Certified Responsible AI Governance and Ethics Professional

KNOWLEDGE CHEATSHEET

CONTROLLED RELEASE

Scope Prioritize → **Test** Gather evidence → **Approve** Accept risk → **Roll out** Observe

- **Cross-validation** evaluates multiple partitions to reduce dependence on a single split. Respect time order, patient or user grouping and other dependencies to prevent leakage.
- **Performance metrics** must match error costs: precision concerns positive predictions; recall concerns actual positives. Inspect subgroup error rates, calibration and relevant thresholds alongside aggregate accuracy or AUC.

Generalization, robustness and change

- **External validation** tests new sites, populations and operational settings. Strong source-site performance does not establish generalization; **transfer learning** adapts learned representations, and adaptation or recalibration requires fresh validation.
- **Edge-case testing** exercises rare but plausible conditions. **Stability testing** checks consistency across relevant variations; load and resilience tests assess behavior under stress or degraded dependencies.
- **Adversarial testing** deliberately manipulates inputs or other attack surfaces to probe resistance. Red-team exercises complement vulnerability scanning and conventional penetration testing.
- **Regression testing** identifies change scope, selects affected and critical scenarios, executes tests and analyzes deviations. New features passing does not prove previously stable behavior remains correct.
- **Accessibility testing** combines automated checks with keyboard, screen-reader and manual evaluation, including users with disabilities. Apply relevant WCAG criteria and test the actual recommendation workflow.

Transparency, explanations and evidence

- **Transparency** communicates purpose, capabilities, limitations and governance. **Interpretability** concerns understanding model behavior; **explainability** provides understandable reasons for outputs. Terminology can overlap.
- **Local explanations** describe an individual result; **global explanations** describe broader model behavior. **Feature importance** estimates input influence, not necessarily causal effect.
- **Explainability interfaces** present decision factors to reviewers and affected users. Test fidelity, stability and audience comprehension; a plausible explanation alone is insufficient.
- **Auditability** requires evidence that can reconstruct decisions and lifecycle changes. Record relevant inputs, model and configuration versions, processing steps, approvals and explanation outputs under privacy controls.

Release gates and controlled rollout

- **Evaluation governance** requires documented validation, fairness, explainability and scenario testing before approval. **Deployment governance** controls promotion, release authority, rollback and operational handover.
- **Model signing** links an artifact digest to a trusted signer. Verify signatures and approved versions during distribution and deployment; hashing alone cannot establish who authorized an artifact.

Testing & Deployment / Lifecycle Assurance

Certified Responsible AI Governance and Ethics Professional

KNOWLEDGE CHEATSHEET

OPERATIONAL ASSURANCE

Monitor Detect → **Contain** Limit harm → **Recover** Validate → **Learn** Improve

- **Canary rollout** serves a small share of real traffic before expansion. **Shadow mode** evaluates a candidate alongside production without letting candidate outputs drive user decisions.
- **Edge AI deployment** supports local, low-latency decisions when connectivity is unreliable. Plan device security, model updates and safe behavior during disconnection.
- **Release decisions** use explicit acceptance criteria and residual-risk approval. Stage expansion when site performance, subgroup fairness or user understanding is inadequate; preserve a safe fallback.

06 Lifecycle assurance and incident response

Monitoring and lifecycle evidence

- **Continuous evaluation** tracks predictive quality, fairness, security, usability and policy alignment in live conditions. **Anomaly detection** flags unusual behavior; alerts need owners and actionable thresholds.
- **Data drift** changes input distributions; **concept drift** changes relationships between inputs and targets. Investigate performance degradation before deciding whether to recalibrate, retrain or change the workflow.
- **Model registries** track versions, ownership, approval and deployment status. **Experiment tracking** records development runs; **metadata and lineage** connect data, transformations, models and historical decisions.
- **Lifecycle assurance** links monitoring, governance review and controlled retraining. Revalidate material changes and affected populations; good uptime does not establish continued trustworthiness.
- **Dashboards** match the audience: executives need strategic exposure, trends and decisions; operators need technical signals and alerts. Use clear thresholds and retain drill-down evidence behind summary indicators.

AI incident triage and containment

- **AI incident playbooks** cover poisoned data, harmful outputs, hallucinations, adversarial inputs and unsafe model behavior as well as infrastructure failures. Preserve relevant data, artifacts, logs and versions.
- **Severity classification** considers actual and potential harm, urgency, reach, autonomy and regulatory exposure. Safety-critical cascading actions demand rapid containment and senior escalation.
- **Containment** may isolate a source, disable an unsafe feature, route decisions to qualified humans or restore an approved model. Restrict throughput only when the remaining operation is demonstrably safe.
- **Circuit breakers** stop high-risk operations when predefined safety conditions are breached. Failover and fallback procedures must preserve safe behavior, not simply service availability.
- **Escalation** moves from first responders through technical and specialist support to management when authority or risk exceeds operational limits. Define titles and thresholds locally; executive decisions include continuity and external notifications.

Lifecycle Assurance

Certified Responsible AI Governance and Ethics Professional

KNOWLEDGE CHEATSHEET

OPERATIONAL ASSURANCE

Monitor Detect → **Contain** Limit harm → **Recover** Validate → **Learn** Improve

Recovery, remediation and assurance

- **Recovery** restores controlled operation, validates behavior and increases traffic cautiously. AI failures can recur as data changes; a discrete fix is not evidence that the root cause remains resolved.
- **MTTD** measures detection delay; **MTTC** measures containment delay; restoration or resolution time measures recovery. Define each clock explicitly because **MTTR** can mean repair, restore, resolve or respond.
- **Phased remediation** addresses multiple findings while preserving safe service: prioritize exposure, improve monitoring and explanations, repair evidence gaps and repeat adversarial and subgroup testing.
- **Assurance dossiers** connect mission objectives, test evidence, limitations, findings, residual risks, remediation owners and retest plans. **Independent assessment** and conformance testing add scrutiny where required or appropriate; certification scope matters.
- **Regulatory change management** tracks developments, assesses applicability and impact, updates policy and controls, implements changes and validates effectiveness. Retain evidence linking obligations to responsible owners.

Retirement and secure exit

- **Offboarding** inventories data, models, credentials, interfaces and dependencies. Revoke access, verify vendor deletion commitments and preserve required records under applicable retention rules.
- **Cryptographic erasure** renders encrypted data unreadable by securely destroying all relevant keys, including recoverable copies. It depends on sound encryption and key management; logical deletion alone is insufficient.
- **Secure wiping** sanitizes supported media; **physical destruction** destroys it. Choose a method suited to the storage technology and control available, then document completion and residual obligations.