AI CONTROLS CATALOGUE

Comprehensive AI Controls Library Across Data, Model, Security, Deployment, Monitoring & Governance

1. DATA CONTROLS

1.1 Data Governance Controls

- DC-01: Data ownership and stewardship defined
- DC-02: Data classification applied (Public/internal/confidential/sensitive)
- DC-03: Data retention policies enforced
- DC-04: Data minimisation applied to all training inputs
- DC-05: Data provenance tracking implemented

1.2 Data Quality Controls

- DC-06: Completeness validation
- DC-07: Accuracy verification
- DC-08: Outlier detection and handling
- DC-09: Duplicate detection and removal
- DC-10: Missing value treatment standards defined

1.3 Data Privacy Controls

- DC-11: PII masking or anonymisation
- DC-12: Consent validity checks
- DC-13: Sensitive attribute protection
- DC-14: Differential privacy optionality
- DC-15: Legal basis for processing validated (PDPA/GDPR)

1.4 Data Bias Controls

- DC-16: Bias detection across protected groups
- DC-17: Balanced dataset evaluation
- DC-18: Label quality audit
- DC-19: Reweighting/rebalancing applied when required
- DC-20: Documentation of known dataset limitations

2. MODEL CONTROLS

2.1 Model Architecture Controls

- MC-01: Model type justified and documented
- MC-02: Feature importance documented
- MC-03: Feature sensitivity analysis conducted
- MC-04: Hyperparameter decisions recorded
- MC-05: Black-box model justification required

2.2 Fairness & Ethical Controls

- MC-06: Fairness metrics computed by demographic groups
- MC-07: Threshold adjustments evaluated for fairness
- MC-08: Unfair bias mitigation strategies applied
- MC-09: Ethical risk scenarios documented
- MC-10: Fairness limitations disclosed

2.3 Explainability Controls

- MC-11: Explainability method selected and justified
- MC-12: SHAP/LIME/Feature Importance evaluated
- MC-13: User-facing explanations documented
- MC-14: Explainability limitations recorded
- MC-15: High-risk decisions require HITL explainability

2.4 Performance & Robustness Controls

- MC-16: Performance metrics defined per use case
- MC-17: Stress testing conducted
- MC-18: Out-of-distribution robustness tested
- MC-19: Model confidence thresholds validated
- MC-20: Periodic model validation scheduled

3. SECURITY CONTROLS

3.1 Input Security Controls

- SC-01: Prompt injection filtering
- SC-02: Adversarial input detection
- SC-03: Sanitisation of user-provided inputs
- SC-04: Context window validation (for LLMs)
- SC-05: Toxic/harmful content blocking

3.2 Training Data Security Controls

- SC-06: Data poisoning detection
- SC-07: Trusted data pipelines enforced
- SC-08: Secure data storage for training sets
- SC-09: Dataset integrity validation
- SC-10: Audit trail for training data changes

3.3 Model Security Controls

- SC-11: Model inversion protection
- SC-12: Membership inference resistance
- SC-13: Model extraction defense (rate limiting, watermarking)
- SC-14: Adversarial robustness via adversarial training
- SC-15: Differential privacy applied where needed

3.4 Deployment Security Controls

- SC-16: API authentication & authorization
- SC-17: Rate limiting for model queries
- SC-18: Logging and monitoring of suspicious behaviour
- SC-19: Encryption in transit and at rest
- SC-20: Access controls for model modification

4. DEPLOYMENT CONTROLS

4.1 Pre-Deployment Controls

- DCY-01: Model Card completed
- DCY-02: Data Card completed
- DCY-03: Risk Assessment completed
- DCY-04: ATLAS Threat Assessment completed
- DCY-05: Governance sign-off obtained

4.2 Release Management Controls

- DCY-06: Versioning strategy enforced
- DCY-07: Code–model consistency ensured
- DCY-08: Shadow mode evaluation (if required)
- DCY-09: A/B testing results documented
- DCY-10: Go-Live checklist completed

4.3 Operational Controls

- DCY-11: Access control matrix defined
- DCY-12: Model fallback strategy (rule-based fallback)
- DCY-13: Manual override capability enabled
- DCY-14: Safe failure mode defined
- DCY-15: Automatic blocking of high-risk outputs

5. MONITORING & AUDIT CONTROLS

5.1 Performance Monitoring Controls

- MON-01: Metrics monitored regularly
- MON-02: Drift detection active (data & concept drift)
- MON-03: Latency & throughput alerts configured
- MON-04: Model degradation thresholds defined
- MON-05: Automatic alerts for anomalies

5.2 Bias & Fairness Monitoring Controls

- MON-06: Fairness metrics monitored by group
- MON-07: Monthly fairness audits performed
- MON-08: Ethical incident log maintained
- MON-09: Fairness degradation triggers retraining
- MON-10: Harmful output escalation workflow

5.3 Security Monitoring Controls

- MON-11: Adversarial detection logs reviewed
- MON-12: Suspicious API usage alerts
- MON-13: Unauthorized access alerts
- MON-14: Attack pattern monitoring
- MON-15: Continuous threat intelligence updates

5.4 Audit Controls

- MON-16: AI Audit Log maintained
- MON-17: Model performance audit trail recorded
- MON-18: Annual external audit required (if high-risk)
- MON-19: Quarterly governance review
- MON-20: Traceability across versions provided

6. GOVERNANCE & OVERSIGHT CONTROLS

6.1 Governance Structure Controls

- GOV-01: RACI defined for all AI roles
- GOV-02: AI Governance Board established
- GOV-03: AI Program Manager oversight
- GOV-04: AI Risk Manager validation
- GOV-05: Cross-functional approval required

6.2 Policy & Compliance Controls

- GOV-06: Alignment with ISO/IEC 42001
- GOV-07: Alignment with NIST AI RMF
- GOV-08: Alignment with OECD Principles
- GOV-09: Compliance with EU AI Act (if applicable)
- GOV-10: Compliance with PDPA/GDPR

6.3 Documentation Controls

- GOV-11: Mandatory documentation package (8 templates)
- GOV-12: Change log for all updates
- GOV-13: Version control enforcement
- GOV-14: Retirement & decommissioning plan
- GOV-15: Annual refresh of all documentation