InfoGovANZ

AI Testing Plan (Template)

1. Purpose and Scope

- **System name:**
- **Business purpose / outcomes:**
- **In scope:** components, data sources, integrations, user groups.
- **Out of scope:** what this plan does not cover (and why).
- **Impact level / risk class:** from AI Screening Tool (normal / elevated).

2. Roles and Responsibilities

- **System Owner:** accountable for approval and operational sign-off.
- **Data Science Lead:** test design, execution, and reporting.
- **Privacy Officer / PIA Owner:** privacy controls, APP alignment, review.
- **Information Security Manager:** security and resilience testing.
- **Risk & Compliance:** independent challenge, evidence review.
- **Internal Audit (if elevated risk):** independent validation.
- **RACI:** attach a one-page RACI for key test activities.

Note: Where no dedicated data-science role exists, testing responsibilities may be undertaken by information management, ICT, or risk or records professionals, supported by external expertise where required.

3. Lifecycle Test Stages

- **3.1 Pre-deployment (lab/QA):** data, model, privacy, security, explainability.
- **3.2 Pilot / shadow mode:** limited production exposure, user feedback, thresholds.
- **3.3 Post-deployment (BAU):** drift, bias, performance, security patching, retrain checks.
- **Entry/Exit criteria** for each stage (what evidence is required to proceed).

4. Test Inventory and Acceptance Criteria

Tailor these to the use case; for elevated-risk systems, include independent testing.

Test Area	What is Tested	Metric / Method	Acceptance Criteria
Data quality	Completeness, accuracy, timeliness, lineage	Profiling; % missing; schema checks	≤1% critical missing, 0 schema errors

Test Area	What is Tested	Metric / Method	Acceptance Criteria
Representativeness	Group coverage vs population	PSI/KS tests; stratified sampling	Each cohort ≥ sample minimum; no severe drift
Performance	Predictive quality	AUC/F1/MAE as relevant	Meets or exceeds baseline (e.g., AUC ≥ 0.82)
Fairness	Outcome and error parity across cohorts	SPD/TPR parity, EO/EOdd	Disparity ≤ 10% or approved mitigation
Robustness	**Sensitivity to small perturbations**	**Stress tests; adversarial prompts (if GenAI)**	**No critical degradation; bounded variance**
Explainability	Local/global explanations	SHAP/LIME + plain-English summaries	Explanations generated and understood by users
Privacy	Purpose compatibility; minimisation; notices	PIA checks; data maps; APP 5/6/11 review	No APP gaps; residual risks documented
Security	Access, encryption, secrets, dependency risk and supply chain risk, AI specific misuse and integrity risks	RBAC tests; pen-test; SCA/SBOM, penetration testing; SCA/SBOM; model integrity and prompt manipulation testing; adversarial or red-team testing where appropriate.	All high findings remediated or formally accepted, including risks arising from adversarial or misuse scenarios.
Usability	Human-in-the-loop flow, escalation	UAT scripts; time-to-decision	≥95% tasks completed within SLA; clear override
Resilience	Failover; rate limiting; timeouts	Chaos tests; load tests	Meets SLOs under peak and failover

Notes:

- SPD = Statistical Parity Difference; EO/EOdd = Equal Opportunity/Equalized Odds; PSI/KS = Population Stability Index/Kolmogorov–Smirnov.
- Techniques may include explainability tools (e.g. SHAP, LIME), role-based access controls (RBAC), prompt testing, and adversarial or misuse testing, depending on system risk.

5. Test Datasets and Control

- **Training / validation / test splits:** proportions and rationale.
- **Holdout policy:** separate, unseen data for final acceptance.
- **Sensitive attributes:** how collected/handled (proxy detection).
- **Synthetic data:** when used and why; limitations noted.
- **Data retention/disposal:** links to records and retention authorities.

6. Procedures

- **6.1 Execution:** who runs each test, tools used, scripts location.
- **6.2 Evidence capture:** where results, logs, dashboards are stored (repo, SharePoint).
- **6.3 Defects & re-test:** severity levels, remediation windows, re-test triggers.
- **6.4 Change control:** model update gates, rollback steps, versioning, SBOM updates.

7. Monitoring Plan (Post-deployment)

- **KPIs:** accuracy, latency, drift scores, fairness deltas, incident counts.
- **Alerts & thresholds:** what triggers amber/red; who is paged.
- **Bias & drift cadence:** e.g., monthly fairness review; weekly drift checks.
- **Human oversight:** which cases require mandatory human review (criteria/thresholds).
- **Reporting:** dashboards to System Owner; quarterly report to AI Governance Committee.

8. Compliance and Alignment

- **APPs:** 5 (notification), 6 (use/disclosure), 10 (quality), 11 (security), 12/13 (access/correction).
- **Standards:** ISO 42001 (Ops & controls), ISO 23894 (risk & monitoring), NIST AI RMF (Manage).
- **Internal links:** AI Policy, PIA, Risk Assessment, Security Testing Report, AI Register entry.

9. Exit / Go-Live Criteria

- All acceptance criteria met or risk-accepted at appropriate level.
- PIA and security findings closed or formally accepted.
- User training completed; SOPs for override and appeals in place.
- AI Register updated; monitoring dashboards live; incident runbook active.

10. Approvals

- **System Owner:** name/date
- **Risk & Compliance:** name/date
- **Privacy (review):** name/date
- **Information Security:** name/date
- **Internal Audit (if applicable):** name/date