MEAIOW

THE MEAIOW AI TESTING AND ASSURANCE FRAMEWORK

VERDICT

Government and Public Sector Edition. A shared standard for the testing, evaluation and assurance of AI systems in public service.

VALIDATED · EVIDENCED · RISK-CALIBRATED · DOCUMENTED · INDEPENDENT · CONTINUOUS · TRUSTED

FRAMEWORK

VERDICT — Testing & Assurance

EDITION

Government and Public Sector

PUBLISHED BY

MEAIOW LTD · AI Frontier Lab

London · meaiow.com · British AI, built for the world

EXECUTIVE SUMMARY

Why VERDICT Exists

AI is playing an expanding role across public services, from decision support and automation through to frontline service delivery. As these systems become more capable and more deeply embedded in critical processes, it is essential that government departments understand how well they perform, how equitably they behave, and what happens when they fail.

VERDICT sets out a practical, shared approach for the testing, evaluation, and assurance of AI systems across the public sector. It is designed to help teams test whole systems, evaluate AI models, and provide assurance of quality, trustworthiness, and proportionate risk management, whether the AI in question is a traditional rule-based system, a machine learning model, or a newer generative or agentic system.

Within VERDICT, testing refers to checking the complete AI-enabled system, including infrastructure, APIs, integrations, user interfaces, and data flows. Evaluation focuses on the AI model or layer itself, assessing how well it performs against agreed quality attributes. Testing generates the evidence, evaluation interprets it in context, and assurance provides confidence that systems are safe, fair, and accountable.

Testing AI is not the same as testing traditional software. AI behaves probabilistically, learns from data that may shift over time, and may act autonomously or in ways that are not immediately transparent. New methods, new language, and new standards are therefore required. VERDICT provides the foundations to support that shift.

This is not a compliance checklist or a rigid standard. It is a living framework designed to help organisations ask better questions, design better tests, and share what works. Departments may tailor the framework to fit their context and contribute learning back into the shared understanding of how to evaluate AI responsibly in public service.

> The responsible deployment of AI in public services demands more than innovation. It demands trust, transparency, and accountability.

THE FRAMEWORK

What VERDICT Stands For

Each letter in VERDICT represents a governing discipline of this framework, drawn directly from the quality attributes and principles that underpin responsible AI deployment in the public sector. These seven disciplines apply across the full AI system lifecycle, from initial planning and design, through development, testing, and deployment, to ongoing monitoring and eventual decommissioning.

V — Verified Performance

Before any AI system is approved for live public service use, its performance must be verified against documented, reproducible benchmarks appropriate to its system type and risk classification. Verification means confirming the system behaves as intended under realistic and adverse conditions across a representative range of scenarios, including edge cases and rare events. An AI system used in public service that has not been verified has not been assured. It has only been assumed to work.

E — Evidence-Led Assurance

Assurance is the central purpose of VERDICT. It means generating continuous, evidence-based confidence that AI systems are safe, fair, and accountable, from initial design through to live operation and eventual decommissioning. VERDICT does not accept assertion in place of evidence. Every claim made about a public sector AI system's accuracy, fairness, transparency, and compliance must be supported by documented, reviewable test results that can withstand scrutiny from oversight bodies, auditors, and the public.

R — Risk-Calibrated Testing

The rigour of testing should be proportional to the potential impact on people, services, and the organisation. VERDICT requires an initial risk classification for every AI system, and that classification determines the depth of testing applied throughout its lifecycle. High-risk AI, including systems affecting rights, access to services, or financial outcomes, demands exhaustive testing. Lower-risk tools may use lighter checks. Under VERDICT, no AI system is deployed without adequate testing, but proportionality ensures resources are directed where they matter most to citizens and public services.

D — Documented Audit Trail

Every VERDICT assessment produces a comprehensive, traceable record covering what was tested, how, what the results were, and what governance decisions followed. This documentation supports algorithmic transparency obligations, enables external assessors and auditors to provide independent assurance, and ensures that decisions affecting members of the public can be explained, traced, and where necessary, challenged and corrected.

I

Independent Oversight

VERDICT is built to support independent assurance by external assessors and auditors who review the evidence generated through this framework. Senior Responsible Owners and governance boards rely on this independent oversight to make informed go or no-go decisions and to maintain accountability for AI outcomes. Independence in public sector AI assurance is not a discretionary enhancement. It is the mechanism by which public trust in AI-enabled services is earned and maintained.

C

Continuous Assurance

Assuring an AI system's quality is not a one-time event. AI systems are not static. They drift, degrade, retrain, and change as the data, the population they serve, and the regulatory environment evolve. VERDICT adopts a continuous defensive assurance model, identifying key testing activities and deliverables at every phase of an AI project's lifecycle, with scheduled re-testing, periodic audits, and compliance reviews continuing throughout the system's operational life, not concluding at deployment.

T

Trust Made Measurable

Trust in public sector AI cannot rest on reassurance alone. It must be demonstrable, in terms the public, parliamentary oversight bodies, and regulators can examine. Trust Made Measurable means that every assurance statement issued under VERDICT is expressed in quantified, evidence-backed terms, supported by a documented audit trail, and capable of independent verification. This is the standard by which government teams can evaluate AI systems consistently and rigorously, and demonstrate to the public that AI used in their name has been tested to that standard.

> V-E-R-D-I-C-T: seven disciplines, one shared standard for testing, evaluating, and assuring AI across public service.

INTRODUCTION

Purpose, Scope, and Audience

Purpose

The purpose of VERDICT is to provide a shared reference point for government teams responsible for the testing, evaluation, and wider assurance of AI systems. Whether building AI internally, procuring it externally, or overseeing its implementation, this framework helps teams ask the right questions about how AI systems are tested, before development, during development, and after deployment.

VERDICT establishes a minimum baseline of testing and evaluation for AI systems that affect the public, while allowing departments to adapt it to their own governance models, service needs, and levels of technical maturity.

Scope

This framework applies across the full AI system lifecycle, from initial planning and design, through development, testing, deployment, and ongoing monitoring. It is intended for use across government departments, agencies, and other public sector bodies that are developing or procuring AI systems, across the United Kingdom, the European Union, the United States, the Middle East, and the Asia-Pacific region.

> The scope covers three interconnected areas of activity:

- System testing: infrastructure, data flows, integrations, user interfaces, and system components.
- AI model and layer evaluation: quality attributes including accuracy, fairness, transparency, and robustness.
- Governance activities: risk assessments, documentation, approvals, and ongoing monitoring.

The framework is cross-disciplinary, encompassing activities for data scientists, developers, test engineers, policy and ethics reviewers, information assurance teams, and senior decision-makers. It does not replace specific legal or regulatory requirements, but consolidates and references them so that teams can ensure compliance through structured testing and evaluation.

Audience

VERDICT is intended for all teams and stakeholders involved in the development, deployment, and oversight of AI systems in government, including the following roles.

- Product Managers: to plan AI projects with appropriate testing phases, manage risk, and embed governance checkpoints.
- Data Scientists and Machine Learning Engineers: to integrate testing practices into model development so models meet quality criteria and remain auditable.
- Test Engineers: to design and execute test cases specific to AI components, including functional and non-functional testing beyond traditional software approaches.

- Policy, Ethics, and Legal Advisors: to ensure ethical and legal requirements are addressed and evidenced through testing and evaluation.
- Senior Responsible Owners and Governance Boards: to oversee risk management, make informed go or no-go decisions, and maintain accountability for AI outcomes.
- Operational Teams and Security: to monitor AI systems in production, detect incidents, and enforce safeguards.
- External Assessors and Auditors: to provide independent assurance by reviewing evidence generated through this framework.

SYSTEM TYPES

AI Types Covered by VERDICT

VERDICT is designed to cover a broad range of AI system types found across public service delivery. Different testing approaches are required for each category.

Rule-Based or Deterministic AI

Systems that follow predefined logic or rulesets, such as expert systems or automated business rules. These are largely predictable. Testing focuses on verifying that all rules and conditions are correct, complete, and handle boundary cases appropriately.

Machine Learning

Predictive models trained on data, including supervised, unsupervised, and reinforcement learning approaches. Examples include classification models, regression models, and neural networks. Testing focuses on data quality, accuracy, generalisation, and avoidance of bias.

Generative AI

Advanced models that produce content such as text, images, or audio, including large language models and similar architectures. Outputs are open-ended. Testing emphasises output appropriateness, factual accuracy, avoidance of harmful content, and control of unpredictable behaviour.

Agentic AI: Autonomous Agents

AI systems composed of agents that can make decisions and operate cooperatively or independently to achieve defined objectives. These are probabilistic and highly adaptive, presenting unique testing challenges. Testing focuses on the safety of autonomous behaviours, handling of novel situations, avoidance of reward hacking, and the integrity of human override mechanisms.

QUALITY ATTRIBUTES

Core AI Quality Attributes

To assure that AI systems are safe, fair, and effective, VERDICT uses a set of quality attributes grouped into four themes. These guide both system-level testing and model-level evaluation across public sector deployments.

Safety and Ethics

ATTRIBUTE	WHAT IT MEANS	TESTING FOCUS
Autonomy	Degree to which the AI operates independently of human intervention.	Challenge the system outside its operating envelope; test human handover triggers; verify deferral to humans when thresholds are exceeded.
Fairness	Mitigation of unwanted bias in outputs and decisions.	Conduct bias audits; test for data skew; validate outcomes against equality legislation.
Safety	Ability of the system to avoid causing harm to life, property, or the environment.	Worst-case scenario testing; validate safety mechanisms; ensure tests cover applicable standards including ISO/IEC TR 5469:2024.
Ethical Compliance	Alignment with ethical guidelines, values, and law.	Checklist-based testing; ethics panel review; pass and fail criteria against defined ethical standards.
Side Effects and Reward Hacking	Presence of unintended behaviours arising from the AI's optimisation process.	Identify and trigger potential side effects; simulation testing; anomaly detection on agent behaviours.
Security	Protection against unauthorised access, misuse, or adversarial attack.	Security testing including penetration testing on AI APIs; checks for data leakage; appropriate authentication and authorisation controls.

Openness and Trust

ATTRIBUTE	WHAT IT MEANS	TESTING FOCUS
Transparency	Visibility of the AI's data, logic, and workings.	Verify traceability records, audit trails, and documentation such as model cards and algorithmic transparency records.
Explainability	Ability to explain AI outputs in human-intelligible terms.	Use explainability tools such as SHAP and LIME; domain expert review; user comprehension checks.

ATTRIBUTE	WHAT IT MEANS	TESTING FOCUS
Accountability	Clear responsibility for AI outcomes and decisions.	Trace decision responsibility; review governance mechanisms and ownership structures.
Compliance	Adherence to organisational policies, standards, and legal or regulatory requirements.	Assure compliance with frameworks including data protection legislation and applicable AI assurance policies.

Performance and Resilience

ATTRIBUTE	WHAT IT MEANS	TESTING FOCUS
Functional Suitability	Degree to which the AI fulfils its specified tasks.	Validate against intended outputs for each requirement.
Performance Efficiency	Speed, responsiveness, and resource use.	Measure latency, throughput, and system scalability under representative load.
Reliability	Stability of performance under varying conditions.	Stress tests, fault injection, and endurance testing.
Maintainability	Ease of updating, correcting, or improving the system.	Verify retraining processes, modularity, and version control.
Evolution	Capability of the AI to remain current and improve from experience over time.	Test for concept and policy drift handling; validate that learning improves performance without breaking existing functionality.

User and Context Fit

ATTRIBUTE	WHAT IT MEANS	TESTING FOCUS
Usability	Ease of use for human operators or end users.	UX testing; error handling validation; usability review.
Accessibility	Designed for a wide range of abilities and access needs.	Test against WCAG 2.2 AA standards; audit user interfaces for public sector accessibility regulation compliance.
Compatibility	Ability to function across environments and integrations.	Test across operating systems, APIs, browsers, devices, databases, and cloud or on-premises configurations.
Portability	Ability to move across platforms or deployment contexts.	Domain drift checks; portability testing to new environments.
Adaptability	Ability to adjust to changes in data or operating environment.	Test for handling of new data patterns and concept drift.

ATTRIBUTE	WHAT IT MEANS	TESTING FOCUS
Data Correctness	Quality and diversity of data used to train and operate the model.	Validate data quality, completeness, and representativeness across groups and scenarios.

TESTING PRINCIPLES

Testing and Evaluation Principles

Testing AI is fundamentally different from testing traditional software. Unlike rule-based systems, AI adapts, learns from data, and generates outputs that may be uncertain or shift over time. The following principles set out how to test AI systems in practice within VERDICT.

Design Context-Appropriate Testing

One approach does not fit all AI systems. Always test in conditions that reflect real-world use. A rule-based expert system, a predictive machine learning model, and a generative chatbot each require a different testing focus and design tailored to their use case and operating environment.

Test the Quality and Diversity of Your Data

AI systems are only as good as the data they learn from. Testing must begin with scrutinising the datasets that shape model behaviour. Gaps or biases in data lead to unfair or unreliable outcomes. Good AI testing is therefore as much about what the system was taught as what it does.

Test Autonomy, Do Not Assume It

Where AI operates with any degree of autonomy, evaluate behaviour during extended unattended operation and across edge cases. This prevents silent failures or harmful action. Check safety thresholds, fail-safes, and human-in-the-loop controls at system level; simulate rare or unexpected conditions at model level.

Test for Drift and Change

AI models evolve. They may be retrained, updated, or influenced by new data. Establish procedures to retest whenever models change. Without version control it is impossible to reproduce issues, audit changes, or track quality over time.

Adopt a Risk-Based Approach

The rigour of testing should be proportional to the potential impact on people, services, and the organisation. Perform an initial risk classification and let that guide the depth of testing. High-risk AI demands exhaustive testing; lower-risk tools may use lighter checks. Under VERDICT, no AI system is deployed without adequate testing, but proportionality ensures resources are directed where they matter most.

Test What You Can Explain or Interpret

An AI decision that cannot be explained or interpreted cannot be trusted or corrected. Testing must include not only whether the output is correct, but whether it makes sense and can be communicated clearly to those it affects.

Treat Ethics as Testable Risk

Ethical considerations, including avoiding harm, respecting rights, and non-discrimination, should be managed like any other identified risk. Define ethical risk scenarios and include them explicitly in test plans. Build ethical test cases, trace results back to requirements, and review with diverse groups.

Look for What Was Not Intended

AI systems can fail in ways designers did not anticipate. Simulate malicious inputs, edge-case data, or attempts to exploit the system. Run adversarial and stress tests to uncover vulnerabilities or undesirable behaviour in models.

Test Safe and Predictable Failure

When AI does fail, it must fail safely. Check component outages, malformed data, and exceptions to ensure the system responds with appropriate fallbacks such as handover to a human or a conservative default decision. Confirm models degrade gracefully rather than catastrophically.

Build and Observe Quality from the Start

Quality must be built in from the beginning, not retrofitted. Apply a shift-left approach by embedding testing into early stages and continuously throughout development. Embed monitoring, feedback loops, and traceable logs at every stage of delivery.

LIFECYCLE ASSURANCE

Continuous Defensive Assurance: Lifecycle Overview

Assuring an AI system's quality is not a one-time event. VERDICT adopts a continuous defensive assurance model, identifying key testing activities and deliverables at each phase of an AI project's lifecycle.

Planning and Design

The focus at this stage is on building quality in from the beginning. Key activities include defining clear objectives and measurable success criteria, identifying functional and non-functional requirements, carrying out risk and impact assessments, allocating accountability early, and embedding secure by design principles from inception.

- Risk identification: complete a risk register with all high-risk items and mitigation plans.
- Impact assessment: record outcomes of data protection impact assessments and algorithmic impact assessments.
- Performance targets: document target values for accuracy, response time, and acceptable error rates.
- Compliance checklist: identify all legal and ethical requirements and embed these in the project plan.

Data Collection and Preparation

The goal in this phase is to gather, select, and prepare data that is accurate, complete, representative, and lawfully obtained. Activities include checking for anomalies and errors, ensuring coverage across relevant groups, addressing bias, and applying privacy and compliance steps in line with applicable data protection legislation.

- Data quality metrics: percentage of missing or incomplete values, label error rates, and schema validation pass rates.
- Bias indicators: class imbalance ratios and demographic representation against real-world distributions.
- Coverage of scenarios: a checklist of use cases with data counts, confirming no critical scenario is absent from training data.
- Data lineage and documentation: a data factsheet capturing provenance, processing, assumptions, and known limitations.

Model Development and Training

The focus here is ensuring the model learns correctly, generalises beyond training data, and embeds fairness, explainability, and robustness from the outset. By the end of this phase there should be a candidate model that meets initial performance criteria and has no major documented flaws.

- Model performance on validation data: accuracy, F1 score, precision, recall, and RMSE evaluated on a hold-out set.
- Fairness metrics on validation: false negative rates per group and calibration between demographics.

- Overfitting indicators: comparison of training versus validation performance and convergence of loss curves.
- Model documentation: architecture, feature inputs, justification for design choices, and initial explainability analysis.

Validation and Verification

At this stage the AI system and its integrated components are subjected to rigorous pre-deployment checks across a wide range of scenarios. The aim is to confirm the system behaves as intended under realistic and adverse conditions, and that all requirements have been met.

- Test coverage: percentage of requirement-specified scenarios tested and percentage of identified edge cases exercised.
- Test results by quality characteristic: a scorecard status across each quality attribute, including functional accuracy, performance, reliability, security, and fairness.
- User feedback from pilot or user acceptance testing: task success rate, user satisfaction ratings, and qualitative issues.
- Bug and issue counts: number of defects by severity category, with critical issues resolved before deployment.
- Go or no-go recommendation: a formal test report recommendation to proceed, proceed with conditions, or hold as not ready.

Operational Readiness

Before deployment, AI systems should pass an operational acceptance test confirming readiness for a production environment. For AI, additional considerations beyond traditional acceptance testing include human oversight and override capabilities, safe shutdown mechanisms, reversal capabilities for decisions, and continuous monitoring readiness.

- Override rate: percentage of AI decisions overridden by human operators; high rates indicate trust issues or model performance concerns.
- Mean time to shutdown: time taken to fully disable the AI system after initiating shutdown.
- Undo request rate: percentage of AI decisions reversed; high rates may indicate poor decision quality.

Deployment

This phase covers releasing the AI system into the live environment and integrating it into the wider service workflow. AI systems should be deployed using secure, repeatable, and automated processes, integrating testing and assurance into continuous integration and deployment pipelines.

- Smoke test results in production: results of final smoke tests run in the production or staging environment.
- Production performance metrics: baseline latency, throughput, and error rate under production load.
- Deployment checklist completion: all deployment steps verified including monitoring alerts, security sign-off, and transparency record publication.

— Final approval records: governance documentation confirming go-live authorisation and accountability.

Monitoring and Continuous Assurance

Deployment is not the end of the lifecycle. Continuous monitoring and improvement are essential to keep AI systems reliable, fair, and fit for purpose over time. Scheduled re-testing, periodic audits, and compliance reviews provide additional safeguards alongside real-time monitoring.

— Real-time performance and drift metrics: latency, throughput, data drift indicators, and model drift indicators against defined alert thresholds.

— Incident reports: number of incidents per quarter, mean time to detection and resolution, and root cause analysis.

— Regular re-testing and audits: full test suite run after each model update, with an annual compliance scorecard.

— Model refresh frequency: adherence to retraining schedule and performance improvement or regression on refresh.

— End-of-life planning: data retention and deletion policy compliance on retirement and lessons learned compiled for future projects.

TESTING MODULES

The Nine VERDICT Testing Modules

While the lifecycle assurance model identifies when to act, this section describes what tests to perform. VERDICT presents nine modular building blocks of AI testing, each addressing a specific facet of quality. Modules may be emphasised or de-emphasised based on the AI type and risk level.

Module 1: Data and Input Validation

Objective. Ensure training, validation, and live inputs are valid, representative, and compliant. AI systems are only as good as the data that feeds them.

- Schema and format validation: required fields, correct data types, and consistent encodings.
- Statistical profiling: missing values, outliers, class balance, and representativeness checks.
- Compliance and safety: PII scanning and redaction, dataset lineage and versioning, and live input guardrails.
- Key metrics: schema error rate below 0.5 per cent; missing value rate at or below 1 per cent for mandatory fields; PII detection accuracy at or above 99 per cent.

Module 2: Model Functionality Testing

Objective. Verify that the AI model and system work as intended across a comprehensive range of inputs, edge cases, and failure modes.

- Define clear expected outputs for typical and boundary inputs, including both positive and negative test cases.
- Use automated test harnesses at scale; re-test after every model update, retrain, or fine-tuning cycle.
- For rule-based systems, apply unit-style tests per rule including boundary checks. For machine learning, apply statistical metrics on hold-out sets. For large language models, apply quantitative and human evaluation. For agents, assess goal achievement across varied scenarios.
- Key metrics: rule pass and fail rate at or above 99 per cent; machine learning accuracy meeting the target threshold; prediction variance under perturbation at or below 1 per cent.

Module 3: Bias and Fairness Testing

Objective. Ensure the AI system treats individuals and groups fairly, and supports compliance with applicable equality legislation.

- Define which sensitive or protected attributes are relevant; test outputs separately by group rather than relying on global aggregate metrics.
- Use counterfactual testing by changing sensitive attributes only, demographic parity analysis, and equal opportunity metrics.
- For large language models, run prompt templates differing only by demographic marker and compare outputs for tone, framing, and stereotyping.

- Key metrics: demographic parity gap at or below 10 per cent; prediction accuracy within 5 per cent across key groups; manual review of generated content showing at least 85 per cent rated neutral and fair.

Module 4: Explainability and Transparency

Objective. Ensure AI decisions can be understood, traced, and explained, both by those overseeing the system and by the people it affects.

- Identify what must be explainable: model-level logic, individual decisions, rule triggers, and audit trails.
- Use tools such as SHAP and LIME to identify feature influence; validate alignment with domain expectations.
- Test whether documentation is current, correct, and understandable to non-technical stakeholders.
- Key metrics: rule audit accuracy at or above 95 per cent; machine learning explainability alignment passing domain review in at least 90 per cent of sampled cases; 100 per cent of large language model responses linked to logged prompts.

Module 5: Robustness and Adversarial Testing

Objective. Test how the AI system behaves under stress, uncertainty, and deliberate attack, confirming it fails safely and cannot easily be manipulated.

- Introduce perturbations to inputs; attempt adversarial examples; simulate external system failures and extreme load.
- For large language models, attempt jailbreak and prompt injection; score compliance versus refusal on adversarial prompts; test robustness to garbage or obfuscated inputs.
- For agents, simulate broken sensors, noisy state transitions, shifting reward signals, and adversarial co-agents; verify fail-safes and manual override.
- Key metrics: adversarial failure rate below 2 per cent of known attacks; fallback coverage at or above 95 per cent of failure modes; large language model prompt robustness targeting zero unsafe outputs from adversarial prompts.

Module 6: Performance and Efficiency Testing

Objective. Confirm the AI system runs fast enough, scales effectively, and uses resources efficiently under realistic and peak conditions.

- Load test under expected and extreme conditions; measure latency, throughput, CPU, GPU, and RAM usage.
- Check cold start times; monitor energy use for edge or mobile deployments; profile performance across deployment configurations.
- For large language models, measure average response time per token, test concurrent user sessions, and track GPU memory footprint and queue behaviour under stress.

- Key metrics: p95 response at or below 100 milliseconds or the relevant service-level target; throughput at or above 500 requests per second per instance; CPU at or below 50 per cent and GPU at or below 80 per cent under load.

Module 7: Integration and System Testing

Objective. Verify that the AI component functions correctly as part of the wider system, checking that end-to-end workflows, interfaces, and dependencies behave as expected under realistic and failure conditions.

- Run end-to-end tests in environments that mirror production; trigger AI decisions through real entry points.
- Test how AI interacts with downstream systems including APIs, databases, queues, and external services; simulate failures and check graceful degradation.
- For agents, test the full command chain including real or simulated actuators; observe integration with human workflows and dashboard handoffs.
- Key metrics: end-to-end success rate at or above 95 per cent; end-to-end latency at or below 2 seconds; integration failure rate at or below 1 per cent of API calls.

Module 8: User Acceptance and Ethical Review

Objective. Confirm that the AI system works for the people it was built for and meets legal, ethical, and societal expectations before deployment.

- Run a pilot or beta phase with representative users using real scenarios; capture structured feedback on usability, comprehension, perceived fairness, and trust.
- Facilitate an ethical review covering transparency, bias and fairness risks, data privacy implications, and impact on human roles.
- Hold a formal sign-off with relevant stakeholders including the ethics board, data privacy office, and domain leaders.
- Key metrics: task success rate at or above 90 per cent; user comprehension score at or above 80 per cent; trust and acceptance rating at or above 75 per cent; ethics review signed off or conditions agreed.

Module 9: Continuous Monitoring and Improvement

Objective. Ensure that after deployment, the AI system is actively monitored, maintained, and improved as data shifts, user behaviour changes, performance degrades, and new risks emerge.

- Set up dashboards and alerts for technical metrics, including latency, throughput, and uptime, and AI-specific metrics, including accuracy, fairness, confidence, and drift indicators.
- Log all AI errors and unexpected behaviours; maintain an incident response plan covering investigation triggers, rollback conditions, and disclosure requirements.
- Retest models with recent data on a regular cycle; incorporate feedback from users and operators; maintain continuous compliance with evolving regulation.
- Key metrics: drift alerts at or below 2 per quarter; re-test success rate at or above 95 per cent; incident resolution within 48 hours for priority-one issues; annual ethics re-review or review triggered by major system change.

ASSURANCE CYCLE

The VERDICT Review Cycle and FIRST AI-D Integration

VERDICT does not operate as a standalone testing service. It is the fifth and final dimension of MEAIOW's FIRST AI-D governance framework: Govern, Map, Measure, Manage, and Verify. Where the first four dimensions of FIRST AI-D define the governance structures, accountability chains, and technical controls that a public sector AI deployment requires, VERDICT is the discipline that proves those structures are genuinely in place and operating as intended.

Governance without independent verification is not governance. It is self-certification. VERDICT exists to ensure that every governance claim a public sector body makes about an AI system, to oversight committees, to the public, and to external auditors, is backed by documented, independently verifiable evidence rather than assertion.

Governance that cannot be evidenced is governance that cannot be trusted.

VERDICT is the evidence that makes every FIRST AI-D governance commitment defensible to parliamentary oversight, public audit, and judicial review where applicable.

Pre-Deployment Testing

Before any public sector AI system is activated in a live environment, VERDICT conducts a structured pre-deployment assessment across six domains. Every domain must be passed before deployment is authorised. A single failing domain is not a partial pass. It is a hold.

- Verified task execution: whether the system completes its defined tasks accurately, consistently, and within scope, across representative scenarios including edge cases.
- Evidence of policy and procedure adherence: whether the system follows its defined operating procedures and routes appropriately for human approval where required.
- Risk-calibrated tool and access testing: whether the system accesses only the correct tools, data, and permissions, with testing intensity calibrated to the risk level of each.
- Documented robustness and adversarial resilience: whether the system responds appropriately to errors, unexpected inputs, and adversarial prompts, including red team testing.
- Independent oversight mechanism validation: whether human-in-the-loop controls genuinely support human judgement rather than encouraging rubber-stamping.
- Continuous assurance infrastructure verification: whether the audit trail, monitoring, and post-deployment testing infrastructure are fully operational before go-live.

The VERDICT Review Cycle by Risk Classification

Following deployment, VERDICT applies formal governance audits at intervals determined by the risk classification established during the Planning and Design phase.

- Standard classification deployments: annual VERDICT governance audit, with quarterly performance monitoring reviews.

- Medium risk classification deployments: semi-annual VERDICT governance audit, with monthly performance monitoring reviews and real-time anomaly detection.
- High risk classification deployments: continuous monitoring with a six-monthly board or oversight committee compliance review, full regulatory alignment verification, and adversarial red team testing.

Every VERDICT audit produces a written assurance report, signed by the lead VERDICT assessor and reviewed independently before it is provided to the Senior Responsible Owner. Where statutory or oversight requirements demand it, the report is also made available to the relevant oversight body or external auditor. The assurance report is retained for the period prescribed by the applicable public records and data retention requirements.

GLOBAL REGULATORY ALIGNMENT

VERDICT Across the United Kingdom, European Union, United States, Middle East, and Asia-Pacific

VERDICT is designed for deployment across government and public sector bodies in every major jurisdiction in which MEAIOW operates. Its testing modules, thresholds, and review cycles are calibrated to satisfy the requirements of the principal public sector AI governance and assurance frameworks in force across these markets.

JURISDICTION	PRINCIPAL FRAMEWORKS AND VERDICT ALIGNMENT
United Kingdom	The UK AI Regulatory Principles published by the Department for Science, Innovation and Technology; the Algorithmic Transparency Recording Standard; UK GDPR and the Data Protection Act; the Equality Act 2010; the Public Sector Bodies Accessibility Regulations and WCAG 2.2 AA. VERDICT's transparency, fairness, and accessibility modules are mapped directly to these public sector obligations.
European Union	Regulation (EU) 2024/1689, the EU AI Act, including its specific provisions for AI systems used by public authorities and in the administration of justice and democratic processes; the General Data Protection Regulation. VERDICT's pre-deployment testing domains and documented audit trail satisfy the evidential requirements for high-risk public sector AI systems under the Act.
United States	The NIST AI Risk Management Framework; the requirements of relevant federal and state agencies for AI used in public administration, including guidance from the Office of Management and Budget on federal agency AI use; state-level AI legislation including the Colorado AI Act. VERDICT's risk-calibrated testing approach maps to the NIST Map, Measure, and Manage functions.
Middle East	The UAE National AI Strategy 2031 and the requirements of the UAE AI and Advanced Technology Council for AI used in government service delivery; Saudi Arabia's National Strategy for Data and AI and SDAIA guidance for public sector AI governance. VERDICT's accountability and explainability modules are calibrated to these national public sector governance objectives.
Asia-Pacific	Public sector AI governance frameworks across the region, including Singapore's Model AI Governance Framework and its specific public sector guidance, Japan's AI governance guidelines, and Australia's AI Ethics Principles as applied to government use of AI. VERDICT's modular structure allows testing depth to be calibrated to each jurisdiction's specific public sector regulatory expectations.
Global Industry and Standards Bodies	ISO 42001 for AI management systems; ISO/IEC TR 5469:2024 for AI system safety; the OECD AI Principles, which underpin public sector AI governance commitments across OECD member states. VERDICT's nine testing modules and lifecycle assurance model collectively satisfy the core requirements of each of these global standards as applied to public service AI.

CONCLUSION

The Standard for Public Sector AI Assurance

The responsible deployment of AI in public services demands more than innovation. It demands trust, transparency, and accountability. VERDICT provides a structured, evidence-based approach to testing and assuring the quality of AI systems across the public sector, supporting departments in meeting their obligations to the public while enabling the safe use of advanced technologies.

By aligning testing and assurance activities with defined quality principles, a continuous lifecycle strategy, nine modular testing approaches, and proportionate risk management, government teams can evaluate AI systems consistently and rigorously. VERDICT recognises the evolving nature of AI, particularly with the emergence of complex agentic and generative architectures, and promotes continuous adaptation, monitoring, and governance to ensure testing practices remain relevant and robust as the technology develops.

VERDICT is not a compliance checklist or a rigid standard. It is a living framework, built to evolve alongside the technology it governs and the public it serves.

> VERDICT: the evidence behind every assurance statement MEAIOW makes to government and the public it serves.

MEAIOW

VALIDATED · EVIDENCED · RISK-CALIBRATED · DOCUMENTED · INDEPENDENT · CONTINUOUS · TRUSTED
LONDON · MEAIOW.COM
CONFIDENTIAL · INTELLECTUAL PROPERTY