IQE Labs
Engineering quality for the AI era

White paper
Version 1.0 · September 2026

AI-Human Quality Assurance Protocol™

A practical framework for trustworthy AI-assisted quality engineering

Trusted
Evidence-based AI-assisted adoption

Controlled
Human guardrails where consequence is highest

Business ready
Reliable releases and measurable value



AI may recommend. Evidence must prove. Humans authorize.

A complete protocol, reference architecture and adoption roadmap to help organizations use AI in testing with confidence, governance and measurable business outcomes.

1 Requirements & risk
2 Test design & data
3 Automation & execution
4 AI analysis & triage
5 Quality assurance
6 Release confidence
7 Real-world outcomes

Executive brief

Trust AI by engineering the controls around it

AI is already entering testing workflows. The enterprise question is whether teams can trust its output enough to act on it.

Large language models are fluent, fast and often useful, but fluent output is not verified truth. A generated test can pass while checking the wrong behavior. A self-healed locator can target the wrong element. A release summary can miss the defect that matters most.

> **The AHQAP starting point**
>
> Assume AI can be wrong, then engineer the controls around it. Trust becomes the product of four properties an organization can build and audit.

Trust =

Evidence	Controls	Accountability	Recoverability
Proves every consequential claim	Constrain what AI is allowed to do	A named human owns each high-risk decision	Every AI action can be traced and reversed

The result is an operating model, not a policy document. AHQAP defines which AI actions need which checks, when a human must approve, how failures are contained, and how autonomy is earned over time. Teams move faster with AI precisely because they no longer rely on it blindly.

In this paper

Why trust fails	The trust problem	03
The protocol	Control loop, lifecycle, autonomy, guardrails, risk matrix	04–08
Controls in depth	Verification, fail-safe design, architecture, release, audit	09–13
Proving it works	Trust metrics, maturity model, applied scenario	14–16
Context	Standards alignment, scope and limitations, about	17–19

AI may recommend. Evidence must prove. Humans authorize.

The principle that anchors every AHQAP control.

© 2026 IQE Labs. Engineering quality for the AI era.

The trust problem

Most AI failures in testing are quiet

The dangerous failure is not always a crash. It is a quality signal that still looks healthy after its meaning has changed.

A model rewrites an assertion so a failing test passes. A triage agent labels a product defect as environment noise. A healing agent changes a locator and the suite turns green. Each step looks reasonable in isolation, while the cumulative result becomes misleading.

The root cause is blind trust: treating fluent output as correct output. AHQAP replaces it with controlled trust, one guardrail for each way AI can quietly degrade the quality signal.

Blind trust: AI without guardrails	Controlled trust: AI with AHQAP
✕ Fluent output treated as truth	✓ Evidence attached to consequential claims
✕ Unclear source or provenance	✓ Approved sources and traceability
✕ Silent test healing	✓ Healing is validated and reversible
✕ Failures can be suppressed	✓ Mandatory failures cannot disappear silently
✕ Unbounded data access	✓ Least privilege and data boundaries
✕ AI judges its own work	✓ Independent verification
✕ Release call on incomplete evidence	✓ Human-owned release decision

Controlled trust beats blind trust.

AHQAP does not make the AI smarter. It makes the use of AI governable. The enterprise requirement is not perfect models, but a system that can constrain, verify, audit and recover from imperfect ones.

© 2026 IQE Labs. Engineering quality for the AI era.

AHQAP at a glance

One eight-control loop for every AI-assisted QE activity

The same control pattern applies whether AI is summarizing logs or influencing a release decision.



1 **Govern**
Define what the AI may and may not do, its risk tier and its owner.

2 **Ground**
Supply only approved, current, access-controlled context.

3 **Generate**
The AI produces a candidate: a test, a label, a fix, a summary.

4 **Verify**
Check the candidate independently with rules and execution evidence.

5 **Approve**
Apply the authority the risk tier demands, up to named sign-off.

6 **Execute**
Act within approved boundaries, versioned and reversible.

7 **Observe**
Watch escapes, overrides, flakiness and customer impact.

8 **Learn**
Feed outcomes back into evaluation sets, thresholds and policy.

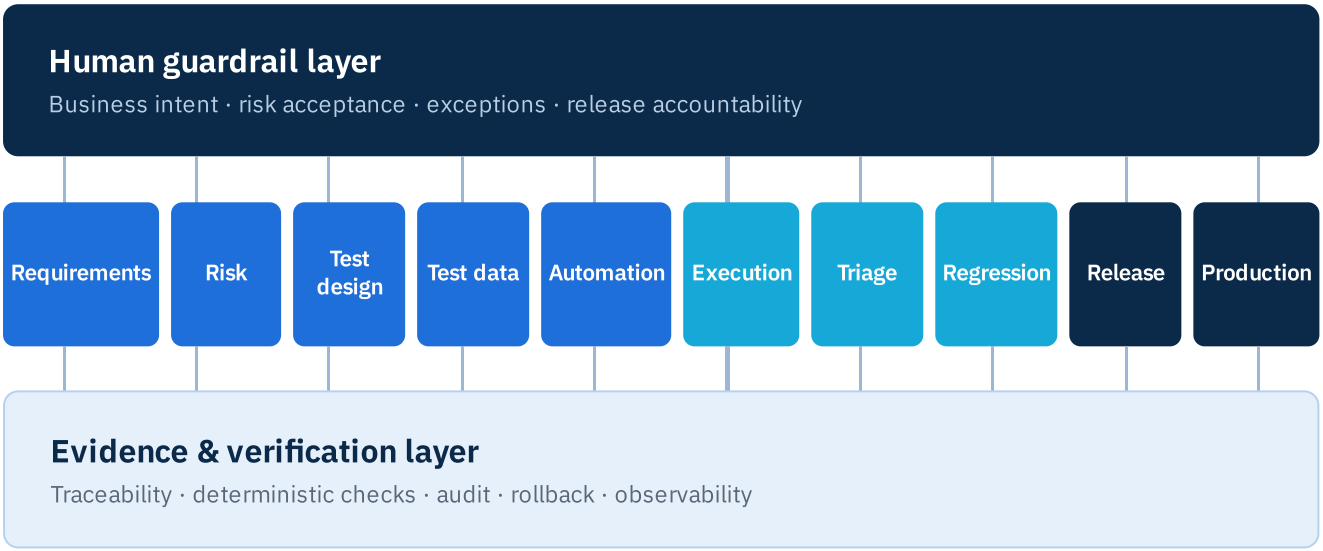
> What the organization learns in **Observe** and **Learn** tightens or relaxes the boundaries set in **Govern**. This is how autonomy is earned rather than assumed.

© 2026 IQE Labs. Engineering quality for the AI era.

End-to-end quality engineering

AI from requirements to production, guarded at every stage

One protocol across the lifecycle, with two constant layers that prevent governance gaps.



What AI does

Analyzes requirements for gaps, generates tests and data, prioritizes regression, executes approved workflows, clusters failures and correlates evidence into recommendations.

What humans retain

Business intent, risk acceptance, high-impact approvals, exceptions and final release accountability: the decisions that should never be delegated.

Where teams usually start

Low risk

Failure triage

Clustering and summarizing failures. Fast value, easy to verify, fully reversible.

Moderate risk

Test generation

New tests reviewed through the normal PR flow, traced to requirements.

Critical, advisory

Release analysis

AI assembles the evidence pack; a named human makes the call.

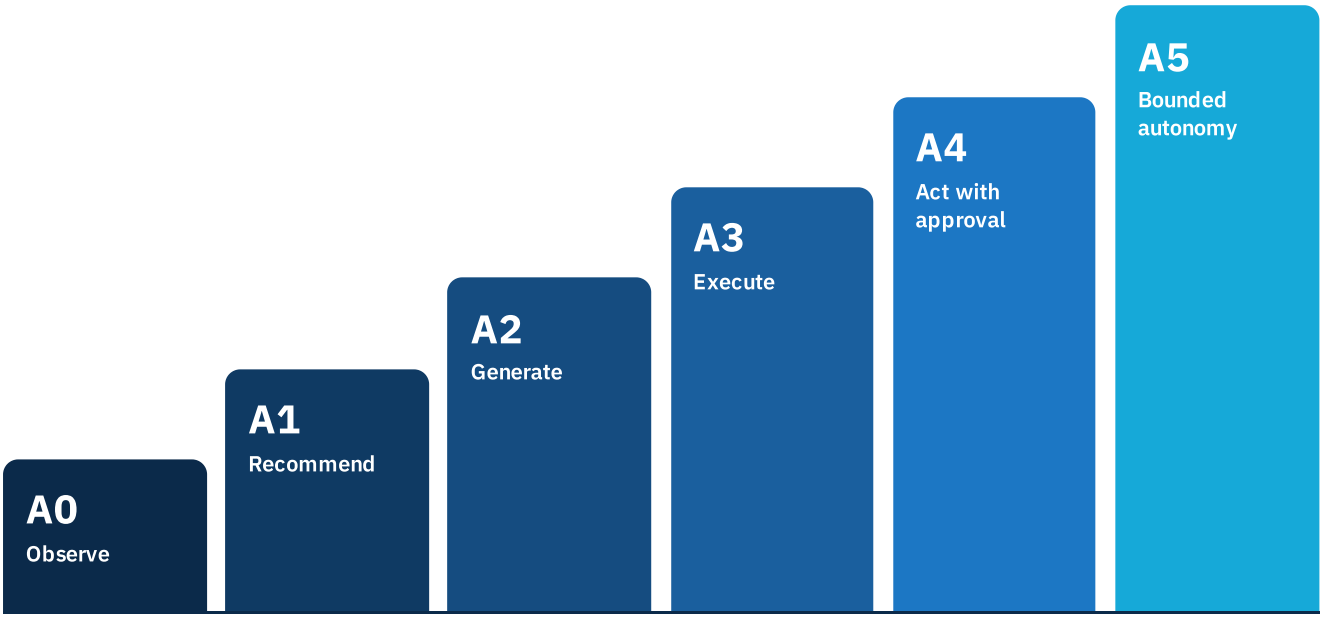
Because the two layers remain constant, a team can start with one use case, such as failure triage, and later expand into test design or release analysis without inventing a new governance model each time.

© 2026 IQE Labs. Engineering quality for the AI era.

Autonomy model

Autonomy is earned by controls, not model capability

A more capable model does not automatically deserve more authority. Each step up requires stronger evidence, better reversibility, closer monitoring and a longer track record.



A0	**Observe**	Read and report only. No suggestions, no changes.
A1	**Recommend**	Suggests actions; a human decides and acts.
A2	**Generate**	Creates tests, data or fixes for human review before use.
A3	**Execute**	Runs approved workflows in controlled environments.
A4	**Act with approval**	Changes shared assets after a named approval per action.
A5	**Bounded autonomy**	Acts without per-action approval, within low-risk, reversible limits.

Boundary rule

A5 is not unrestricted autonomy. It is reserved for low-risk, observable and reversible actions. Release decisions never reach A5.

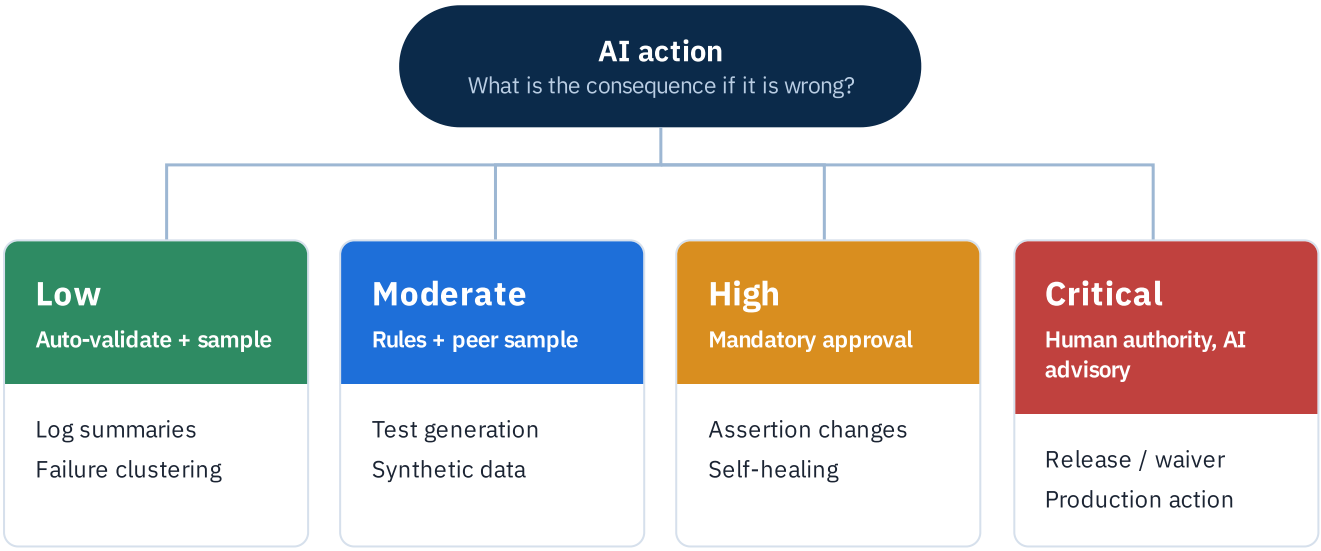
© 2026 IQE Labs. Engineering quality for the AI era.

Human guardrail

Do not force humans into every loop

Uniform oversight creates reviewer fatigue. AHQAP puts human judgment where the consequence of being wrong is highest.

When every AI suggestion needs sign-off, reviewers are flooded, approvals become rubber stamps, and the one review that mattered is lost in the noise. Uniform oversight is weak oversight.



Every AI action is classified by what happens if it is wrong. Low-risk actions are validated automatically and sampled. Moderate actions add deterministic checks and peer sampling. High actions require a named approval. Critical decisions remain under human authority, with AI advisory only.

The objective is not less human oversight. It is better human oversight.

Reviewers spend attention on decisions that can hurt customers, and auditors can see who approved an action and why that level of review was sufficient.

© 2026 IQE Labs. Engineering quality for the AI era.

Risk × autonomy matrix

Risk sets the ceiling. Evidence earns the level.

The consequence of an error caps how much autonomy AI may ever hold, regardless of model confidence.

Risk tier	Max autonomy	Verification	Human role	Typical use
Low	**A5 bounded**	Automated validation + periodic sampling	Reviews samples and trends	Summaries, clustering, flaky-test tags
Moderate	**A3 controlled**	Deterministic checks + peer sample	Normal PR / review flow	Test generation, synthetic data, prioritization
High	**A4 with approval**	Independent verification + intent proof	Named approver per change	Assertions, healing, test deletion or skip
Critical	**A1 recommend only**	Full quality evidence pack	Human decides; AI advisory	Go / no-go, waivers, production

The ceiling is fixed by consequence, not confidence. A model that is right 99% of the time on release recommendations still stays at A1, because the 1% it gets wrong reaches customers. A use case usually starts below its ceiling and moves up as its trust metrics prove out.

1. **Highest tier wins**
   Classify by the most consequential thing the workflow can touch. A test generator that can also edit assertions is governed as High.
2. **Promotion is evidence-based**
   Move up one level only after sustained performance on trust metrics, against organization-defined thresholds.
3. **Demotion is automatic**
   Any fail-safe trigger drops autonomy one level until the issue is corrected and re-verified.

Non-negotiable

A5 exists only in the Low tier. Nothing that can change what a test proves, or what reaches production, runs without accountable human authority.

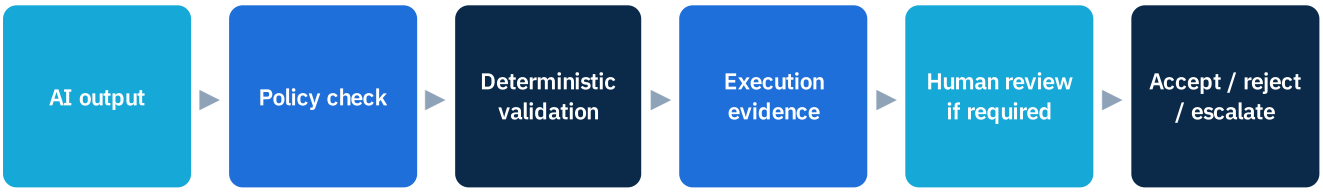
© 2026 IQE Labs. Engineering quality for the AI era.

Independent verification

AI output is a candidate until independently verified

Asking the generator to be the only judge of its own work preserves the same blind spot.

Verification is a pipeline. A policy check first confirms the action is permitted for the use case and risk tier. Deterministic validation then applies schemas, rules, static analysis and traceability. Execution evidence tests the output against known behavior, and human review is added where the risk tier requires it.



Example: generated test

1. Trace it to a requirement
2. Validate data and schema
3. Execute against known behavior
4. Compare expected with actual
5. Human review of critical scenarios

Example: self-healing

1. Prove the locator targets the intended business element
2. Confirm no assertion changed
3. Run the healed test
4. Version the change
5. Retain a rollback path

Anti-pattern

“AI says the test is fixed, therefore the test is correct.”

A passing test proves execution, not preservation of business intent.

© 2026 IQE Labs. Engineering quality for the AI era.

Fail-safe design

When AI is wrong, reduce autonomy

The default response to uncertainty is to stop consequential action, preserve evidence and recover safely.

Every AI system will eventually be wrong. What matters is what happens next. In an unprotected pipeline, an AI mistake degrades quality silently until a customer finds it. In AHQAP, a failed control stops action rather than quietly lowering quality.



Fail-safe triggers, defined in advance

Nobody has to judge in the moment whether something is bad enough.

Evidence missing	Sources conflict	Deterministic validation fails	Policy violated
Unusual model behavior	Permission boundary exceeded	Mandatory test skipped	AI service unavailable

Design goal

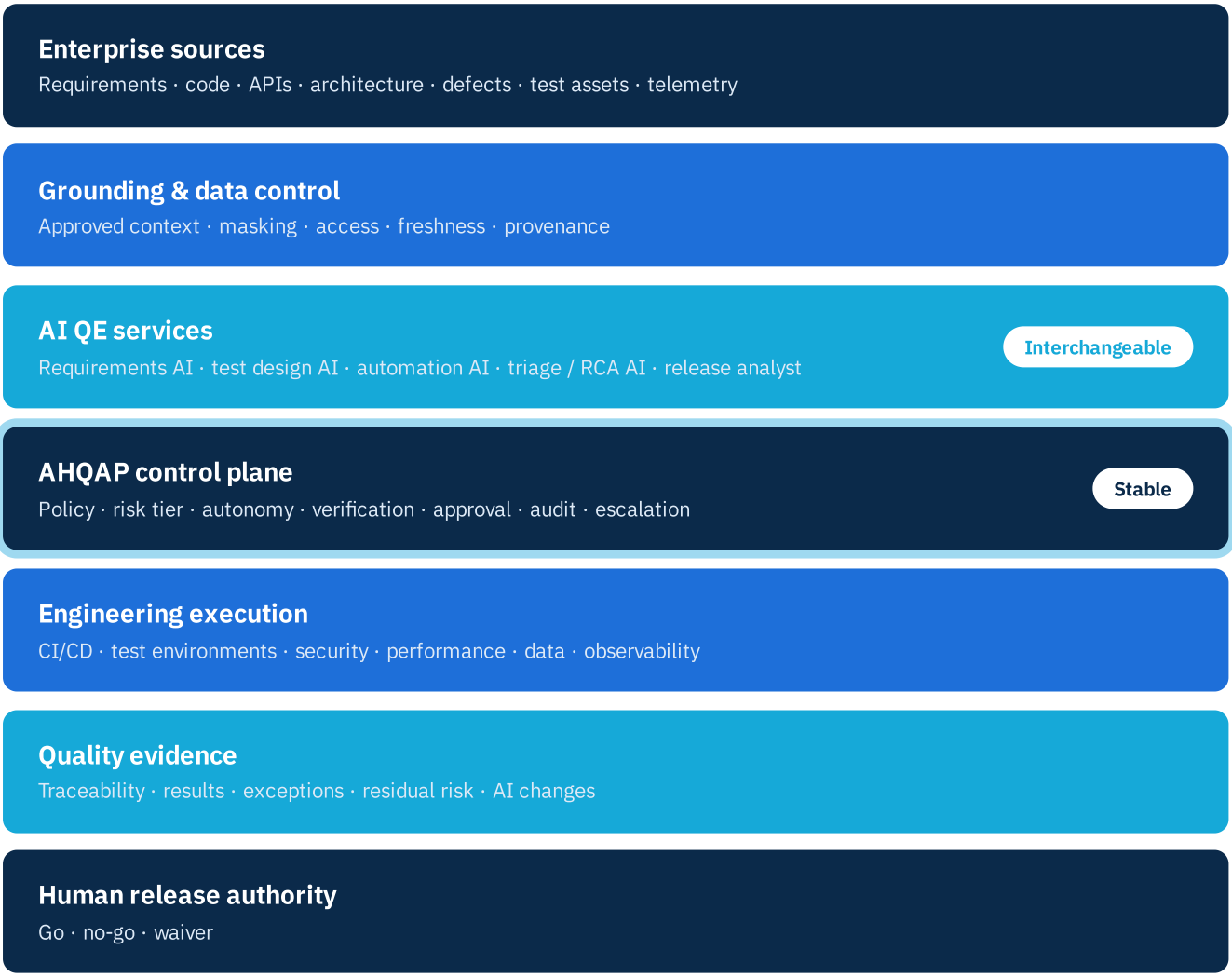
An AI mistake becomes a controlled engineering event with a record, an owner and a recovery path. It does not become a silent customer-facing failure.

© 2026 IQE Labs. Engineering quality for the AI era.

Reference architecture

Keep the control plane stable while AI tools change

Vendor independence comes from separating interchangeable AI capabilities from durable policy, verification and release controls.



Every AI action passes through the control plane. Models and vendors can change every quarter; the rules about what they may do, how they are checked and who signs off do not.

A team can replace one test-generation model with another without rewriting its approval rules, audit trail, evidence requirements or release criteria.

© 2026 IQE Labs. Engineering quality for the AI era.

Release assurance

Evidence before opinion

Release readiness is assembled from evidence first. AI can interpret the pack, but a named human owns the decision.

Release meetings often run on opinion: “it feels stable,” “the suite is mostly green.” AI can make this worse by producing a confident summary that hides gaps in the underlying evidence. AHQAP reverses the order.



The AHQAP difference is visibility

The last two layers make the pack AHQAP-specific: reviewers can see how much of the evidence was touched by AI, and where humans overrode it.

© 2026 IQE Labs. Engineering quality for the AI era.

Auditability & security

Every consequential AI action leaves a trust record

The team must be able to reconstruct what the AI knew, did, validated and changed.

	Record field	Question it answers
1	**Agent / use case**	Who acted
2	**Model + workflow version**	What logic and version acted
3	**Input sources**	What evidence was available
4	**Risk tier + autonomy**	Why the action was allowed
5	**Validation evidence**	What independently passed
6	**Human approval / override**	Who accepted or changed it
7	**Outcome / rollback**	What happened next
8	**Post-release signal**	What later evidence showed

Engineers

Use the record to debug what the AI did and why.

Auditors

Use it to confirm the controls were followed.

The protocol

Uses it to decide whether a use case has earned more autonomy.

Security and privacy baseline

These sit underneath every control.

- ✓ Classify and minimize data
- ✓ Keep secrets out of prompts
- ✓ Define retention and residency boundaries
- ✓ Prefer synthetic or masked test data
- ✓ Enforce least privilege
- ✓ Separate test and production permissions

© 2026 IQE Labs. Engineering quality for the AI era.

Trust measurement

Measure trust. Do not declare it.

A model confidence score is not an operational trust score. AHQAP measures outcomes by use case and risk tier.

A confidence score measures how sure the model is, not how often it is right, and it ignores what is at stake. Eight operational metrics form the core scorecard; each shows its target direction.

Unsupported output rate
Claims without evidence
Falling ↓

AI artifact acceptance
Validated output adopted
Rising ↑

False positive / negative rate
Triage and classification quality
Falling ↓

RCA confirmation rate
Root causes confirmed by engineers
Rising ↑

Human override rate
Autonomy calibration
Stable band ↔

AI-caused escapes
Customer-impacting AI error
Target: zero 0

Critical coverage preservation
Protected business paths tested
Target: 100% 100%

Time to triage
Speed of diagnosis
Falling ↓

Read the override rate carefully

Near zero can mean reviewers stopped looking; very high can mean the AI is not ready for its autonomy level. The goal is a healthy, monitored band. These metrics drive promotion and demotion in the risk × autonomy matrix.

© 2026 IQE Labs. Engineering quality for the AI era.

Adoption path

The IQE Labs maturity model

Moving up is not simply a technology upgrade. It is a controls upgrade.



Your first 90 days

A controls upgrade can start within a quarter.

Days 0–30

Control

Inventory AI use · classify risk · define prohibited actions · baseline metrics

Days 31–60

Verify

Add deterministic checks · trust records · approvals · evaluation datasets · CI/CD controls

Days 61–90

Scale

Expand proven use cases · calibrate autonomy · integrate release evidence · review outcomes

> By day 90, the organization knows where AI is used, what it is allowed to do, and whether it is working.

© 2026 IQE Labs. Engineering quality for the AI era.

Illustrative practitioner scenario

When 80% of a Selenium suite fails

How AHQAP prevents an AI repair agent from turning “green tests” into false confidence.

Overnight, 80% of a large Selenium regression suite fails. An AI repair agent is available and release pressure is high. The tempting move is to let the agent fix everything until the dashboard turns green. That is exactly how a real product defect gets repaired out of sight.

1 **Cluster**
Signature, page, exception, recent change

2 **Classify**
Product / automation / environment / data / unknown

3 **Propose**
Repair only where evidence supports it

4 **Isolate**
Branch or controlled environment

5 **Verify**
Intent and assertions preserved

6 **Approve**
Human validates ambiguity or high risk

7 **Merge**
Normal PR, CI and rollback controls

8 **Report**
Product risk separate from automation health

Without AHQAP
The agent rewrites locators and assertions until the suite passes. The dashboard shows green, and genuine product defects ship hidden inside “repaired” tests.

With AHQAP
Repairs happen only where evidence supports them, in isolation, with intent verified. Product defects surface in their own report instead of disappearing.

AI accelerates diagnosis without equating “tests repaired” with “product safe.”

© 2026 IQE Labs. Engineering quality for the AI era.

Standards alignment

Built to fit the frameworks enterprises already answer to

AHQAP is a quality engineering protocol, not a new compliance regime. It operationalizes governance principles inside day-to-day testing.

Framework	What it asks for	Where AHQAP contributes
NIST AI RMF 1.0	Govern, Map, Measure, Manage; continuous risk management and independent testing and measurement.	Govern and Ground; risk tiers; trust metrics; Verify; fail-safe; Learn. *NIST AI RMF 1.0 is currently being revised.*
ISO/IEC 42001:2023	Establish, implement, maintain and continually improve an AI management system.	Control-plane policy; trust records; risk treatment; 30/60/90 adoption; Learn. *AHQAP can support an AIMS but does not certify conformity.*
EU AI Act	For applicable high-risk AI: risk management, logging and traceability, transparency, human oversight, accuracy, robustness and cybersecurity.	Risk-based human guardrail; trust record; quality evidence pack; human authority; verification. *Applicability depends on the system, use case and legal context.*

For organizations already working toward these frameworks, AHQAP supplies the testing-specific evidence their auditors will ask for. For those that are not yet, it builds the habits those frameworks expect.

Important

AHQAP supports alignment with these frameworks but does not by itself establish or certify legal or standards compliance.

© 2026 IQE Labs. Engineering quality for the AI era.

Scope & limitations

What AHQAP does not claim

Credibility comes from defining the boundary of the framework as clearly as its promise.

No zero-error promise

AHQAP does not guarantee that AI will never hallucinate, misclassify or generate an incorrect artifact.

No zero-defect promise

It does not guarantee defect-free software or risk-free releases.

No compliance certificate

It supports evidence and operating controls, but is not a substitute for legal advice, certification or organization-specific compliance work.

No universal threshold

Promotion thresholds, sampling rates and acceptable risk must be defined for the organization and use case.

No replacement for accountability

High-consequence business and release decisions remain owned by authorized people.

The defensible promise

AHQAP provides a systematic way to constrain AI autonomy, independently verify consequential outputs, preserve human accountability, and create auditable quality evidence.

© 2026 IQE Labs. Engineering quality for the AI era.

About

Practitioner-led quality engineering



AHQAP is designed from the perspective of teams that must ship software, explain risk and stand behind release decisions.

About the author

Kamal Kumar Natha is a quality engineering and test automation leader with more than 21 years of experience spanning automation architecture, QA leadership and enterprise quality transformation across airline, retail, hospitality, banking, insurance and digital commerce.

He is the author of *Beyond Automation: The Future of Testing in the Age of AI* and has developed practitioner work on intelligent and agentic quality engineering. AHQAP brings those ideas into a concrete operating protocol for AI-assisted testing.

About IQE Labs

IQE Labs develops practical approaches for intelligent quality engineering, AI-assisted quality engineering, agentic QE and enterprise quality transformation.

Its focus is on the controls, evidence and operating models that help organizations use AI without shifting unmanaged risk onto customers.

21+
Years in quality engineering and test automation

QE + AI
Automation, data and AI: practitioner-led quality engineering

8
Controls in one protocol, from prompt to production

Our view

Quality is engineered, not tested in.

AI should increase engineering leverage without weakening the evidence behind a release decision.

© 2026 IQE Labs. Engineering quality for the AI era.

Next step

Find out where your AI stands

Most teams already use AI somewhere in testing. Fewer can say which use cases are safe to scale and which are quietly adding risk.

IQE Labs AI Trust Readiness Assessment

A structured assessment that gives you:

- An inventory of current AI use in quality engineering
- Each use case classified by risk tier
- Your position on the IQE Labs maturity model
- Baseline trust metrics
- A prioritized 30/60/90-day control roadmap

Start the conversation

iqelabs.ca

Book a discovery conversation about your AI-assisted testing practice.

Scan to visit

AI Trust Readiness conversation

iqelabs.ca

AI may recommend. Evidence must prove. Humans authorize.

© 2026 IQE Labs. Engineering quality for the AI era.