

From AI Pilots To Governed Scale

Why regulated organisations need an operating model before another proof of concept

Why regulated organisations need an operating model before another proof of concept

The short version

AI pilots prove that a use case might work.

They do not prove that the organisation can run it safely, maintain it, explain it, or defend it.

Regulated organisations need a scaling layer between experiments and live service. That layer should be light enough for delivery teams to use, but strong enough to answer regulatory, customer, operational, data, supplier, and board-level questions.

This paper describes what that layer looks like and how to build it in 30 days.

Introduction

AI pilots are useful. They are also a poor test of whether an organisation is ready to run AI properly.

A pilot can work with friendly users, narrow data, manual checks, temporary permissions, and a team watching every step. That tells you something, but not enough.

It does not prove the use case can survive live operations. It does not prove the data is fit for the decision. It does not prove the supplier can be controlled. It does not prove a human can challenge the output in time. It does not prove the organisation can explain the decision to a regulator, board, customer, auditor, or court.

That is the gap regulated organisations now need to close.

The UK market has moved past the first wave of "can we do something with AI?" The sharper question is now: can this be scaled, governed, maintained, and defended?

The public signals all point the same way. GDS has described the shift from many AI pilots to fewer, larger mission programmes, with scale blockers around value evidence, procurement, commercial constraints, data readiness, AI assurance, safety, legal questions, and adoption planning. GDS and The National Archives have also put data maturity at the centre of AI readiness: managed data, understood data, validation, people, process, and culture.

ICO is turning agentic AI into practical work on data protection, automated decision-making, profiling, cross-regulatory implications, and sandbox support. UK policy is backing assurance and controlled testing rather than pretending one generic rulebook will answer every sector question. In water, AMP8 innovation funding is pulling AI and data into operational outcomes, from catchments and assets to customers and control rooms. Utility tenders are already asking for the real thing: controlled deployment, internal data boundaries, explainability, responsible AI, supplier discipline, and upskilling.

That is the work.

The pilot is not the destination. It is evidence for the next decision.

This paper sets out the scaling layer regulated organisations need between experiment and live service: intake, classification, ownership, evidence, architecture review, supplier checks, release gates, monitoring, and a register that records the decisions people actually made.

Not theatre. Not a 90-page policy that everyone ignores.

Just enough governance to let good AI move, weak AI improve, and bad AI stop before it becomes operational debt.

Why pilots are misleading

A pilot can avoid the hard questions. That is partly why pilots succeed.

Limited data means no one has to explain what happens when the data is incomplete, wrong, or out of date. Friendly users means no one is challenging the output. Narrow scope means edge cases have not appeared yet. Manual workarounds are in place but not counted as controls. The team watching every step cannot scale. There is no live operating model, no customer or regulatory pressure, no supplier-change problem, and no long-term ownership question.

All of that changes when the use case moves to production.

Live data is messier. Users are not all technical or well-disposed. Scope expands. Workarounds become debt. The pilot team moves on. The supplier updates the model. A regulator asks a question nobody thought to anticipate.

The pilot proved that the idea works. It did not prove that the organisation is ready to run it.

That is not a reason to avoid pilots. It is a reason to treat them as learning, not as proof of readiness.

The 7 scale blockers

Most AI use cases in regulated organisations stall, or quietly drift into uncontrolled operation, because of one or more of these seven problems.

1. Value evidence

The benefit is interesting, but not specific enough to justify control cost. Someone believes the model saves time or reduces errors, but there is no agreed baseline, no measurement plan, and no threshold that would justify the oversight burden. Without it, the business case cannot hold up to a board, procurement, or regulator.

2. Data readiness

The data exists, but ownership, quality, provenance, access, retention, and validation are not clear. The use case was built on a snapshot that may not reflect live conditions. Nobody has confirmed that the data is fit for the specific decision the model is being asked to make.

3. Supplier control

Third-party AI is being used without enough visibility of model behaviour, data use, change cadence, or exit route. The supplier may update the model without notice. Data may leave the boundary the organisation assumed it was inside. There is no contractual or technical mechanism to respond to a change.

4. Architecture fit

The use case bypasses integration, security, monitoring, operational resilience, or change-control reality. It was built as a standalone thing and was never connected to the systems it would need to rely on in production. When it hits live infrastructure, the gaps become incidents.

5. Human oversight

A human is named as the accountable party, but they are not trained, informed, or able to challenge the output in practice. They may be approving decisions too quickly to review them, or they may have no mechanism to escalate or reverse an output once it has triggered an action.

6. Evidence design

Approvals depend on confidence rather than recorded evidence. The organisation cannot produce, for a given decision, the data that was used, the rule or model that made it, who reviewed it, and what they checked. That is not a compliance problem in isolation. It is an operational and reputational risk.

7. Operating ownership

Nobody owns the use case after the pilot team moves on. No one is responsible for monitoring, retraining, data drift, supplier changes, user feedback, incident response, or the decision to stop or simplify. The use case runs until something breaks, and then nobody is sure who is responsible for fixing it.

The scaling layer

The scaling layer is not a committee. It is not a new governance framework that delivery teams have to read before they can start anything. It is a set of lightweight, connected processes that answer the seven blockers before they become problems.

Minimum viable model:

Intake. One route for AI use cases to enter the governance process. A short form, not a long one. The purpose is to make every use case visible so it can be classified and owned.

Classification. A simple risk-tiering scheme that routes use cases to the right level of scrutiny. The dimensions that matter: impact on people, degree of autonomy, data sensitivity, affected parties, supplier involvement, and reversibility of outputs. Most use cases are low risk and should move quickly. A few need more.

Ownership. Named business, technology, data, risk, and operational owners before the design choices harden. Not a committee. Named individuals who can make decisions and be held to account.

Architecture review. Before the technical shape of the use case is fixed. The purpose is to catch integration, data boundary, security, supplier, and monitoring gaps while they are still cheap to fix.

Supplier assurance. Before contract or renewal. The minimum: what the supplier does with data, how the model changes, what notice is given, and how the organisation exits if the relationship ends badly.

Evidence pack. Before release. A short record of what was decided, by whom, on what basis, what was checked, and what monitoring is in place. Not a compliance document written after the fact. A real decision record.

Live register. A single list of all AI use cases, their status, risk tier, owners, and next review. Updated when something changes.

Monitoring and change review. A defined process for catching data drift, model changes, supplier updates, and user feedback after release. Including a process for deciding when to stop or simplify.

None of this needs to be heavy. The point is that the organisation has a consistent, legible process that delivery teams actually use, and that the evidence exists if it is ever needed.

Data maturity before AI readiness

AI readiness is not mainly a model-selection problem.

Before a use case can scale, the organisation needs to know:

- what data is being used
- who owns it
- whether it is fit for the decision
- where it came from
- how it changes over time
- who can access it
- what outputs need validation
- what happens when the data or model drifts

If those answers are weak, the AI use case is not ready. It may still be worth exploring. But it should not be treated as ready to scale.

GDS and The National Archives put this plainly in their 2026 data maturity discovery: the foundation for AI-ready government data is managed data, understood data, validation, people, process, and culture. Not more models.

The practical implication is that a data gap found during AI intake is not a reason to stop the use case. It is a reason to fix the data first, or to build the validation into the design. Both are cheaper earlier than later.

Agentic AI raises the control bar

Agentic AI is different because the system can take actions, not just produce outputs.

It can call tools, trigger workflows, send communications, update records, or make decisions that feel operational rather than advisory. The human may be several steps back in the process. By the time the output lands, the action has already happened.

That changes the control questions:

- what can it do
- what can it access
- what can it change
- when must a human approve the next step
- what is logged at each action
- how is harm reversed
- who handles complaints or challenge
- how are permissions reviewed as the system changes

The ICO's 2026 tech futures report on agentic AI addresses this directly: automated decision-making, profiling, data protection, cross-regulator implications, and sandbox routes are all in scope when the system acts on behalf of an organisation.

The practical test is straightforward:

If the system can act, the control model must know what action means.

That includes knowing what actions the system cannot take without human approval, what the log looks like for every step it does take, and who is responsible when something goes wrong.

What water and utilities show

AMP8 is pushing data and AI into operational outcomes: network performance, water quality, customer behaviour, asset intelligence, predictive maintenance, control-room decisions, and environmental monitoring.

This is exactly where weak AI governance becomes expensive.

Yorkshire Water's AI services framework and Thames Water's sewer mapping contract are not the end of the trend. They are the early examples. As more operational decisions are influenced or made by AI systems, the regulatory, operational, and reputational exposure for getting it wrong increases.

In a regulated utility, the question is not just "does the model work?" It is:

- can operations trust it
- can it be explained to Ofwat, the EA, or a public inquiry
- can data stay inside the right operational boundary
- can suppliers be controlled through procurement, contract, and change management
- can the organisation evidence why the use case moved from pilot to live
- can the use case survive audit, incident review, or public scrutiny

Ofwat's innovation funding creates real pressure to move AI from lab to live. The scaling layer is what makes that movement safe.

The proceed, simplify, stop test

Every AI use case review should end with one of three decisions.

Proceed when:

- the use case is specific enough to be governed
- the benefit is measurable against a baseline
- ownership is named and accepted
- data is understood, accessible, and fit for purpose
- supplier risk is visible and contractually covered
- human oversight is real, not nominal

- release evidence is good enough to answer a regulator
- live monitoring is defined before release

Simplify when:

- the use case tries to do too much at once
- deterministic rules would solve most of the problem more safely
- the model is compensating for a process that should be fixed
- the data needs to be improved before the model can work reliably
- the control cost is higher than the benefit at the current scale

Stop when:

- no one owns the outcome
- affected parties cannot challenge or recover from harm
- the organisation cannot explain the decision in plain terms
- access or data boundaries are unclear or contested
- assurance depends on optimism rather than evidence

The test is not designed to block AI. It is designed to make the decision explicit, rather than letting weak use cases drift into production by default.

30-day action plan

This does not need a programme. It needs four weeks of focused work.

Week 1: Map what exists

List every current AI pilot and every supplier product that uses AI. Agree a short intake form. Pick three or four risk tiers. Name an accountable owner for the process.

Week 2: Run the model on live cases

Take five to ten current use cases through the classification, ownership, data readiness, and supplier assurance checks. Identify what is missing. Classify each as proceed, simplify, stop, or wait.

Week 3: Connect it to real gates

Map AI review into architecture, procurement, delivery, risk, security, and release processes. Define the evidence pack by risk tier. Build the first live register.

Week 4: Brief and begin

Brief delivery teams on the intake and classification process. Set the monthly review rhythm. Assign owners for fixes on live or near-live use cases. Pick the first pilot to take through a formal scale review.

The aim at the end of 30 days is not a perfect governance framework. It is a functioning process, a live register, and at least one use case that has gone through the full model and produced a clear decision.

That is enough to start. The process improves as cases go through it.

What an independent review can add

An external reviewer can do three things an internal team cannot easily do: read the whole picture without departmental bias, benchmark current controls against what regulators and buyers are actually asking for, and bring a clean view of where the gaps are before they become incidents.

Occamly's 10-day AI Governance and Architecture Triage looks at three to five use cases, current governance, data and supplier assumptions, architecture fit, and release evidence.

The output is a practical decision pack:

- what can proceed
- what should simplify
- what should stop
- what evidence is missing
- what the next 90 days should fix

It is not a long report. It is a working document with clear recommendations and the evidence to back them.

If AI pilots are moving faster than the controls around them, that is a solvable problem. It is much cheaper to solve it before a live incident than after.

Sources

- GDS AI Exemplars programme, 11 June 2026: <https://gds.blog.gov.uk/2026/06/11/what-we-learned-from-the-prime-ministers-ai-exemplars-programme-and-how-it-is-shaping-the-next-phase/>
- GDS and The National Archives data maturity discovery, 4 June 2026: <https://gds.blog.gov.uk/2026/06/04/data-maturity-the-foundation-for-ai-ready-public-sector-data/>

- Government guidance on making datasets ready for AI:
<https://www.gov.uk/government/publications/making-government-datasets-ready-for-ai/guidelines-and-best-practices-for-making-government-datasets-ready-for-ai>
- ICO agentic AI futures report: <https://ico.org.uk/about-the-ico/research-reports-impact-and-evaluation/research-and-reports/technology-and-innovation/tech-horizons-and-ico-tech-futures/ico-tech-futures-agentic-ai/>
- DSIT AI Adoption Summit speech, 8 June 2026:
<https://www.gov.uk/government/speeches/liz-kendalls-speech-at-ai-adoption-summit>
- Ofwat innovation in the water sector: <https://www.ofwat.gov.uk/regulated-companies/innovation-in-the-water-sector/>
- Yorkshire Water AI Services Framework award notice: <https://www.find-tender.service.gov.uk/Notice/079100-2025>
- Thames Water Sewer Mapping award: <https://www.find-tender.service.gov.uk/procurement/ocds-h6vhtk-0597b6>