Insight ON: AI

Who's Running Your AI?

What 'End-to-End' Actually Means

InsightAI

There's a moment many technology leaders will recognise: their delivery engagement closes, the new system goes live, and their partner moves on to the next project. From then on, the organisation is responsible for something it didn't fully design for long-term operations: governance frameworks that were always going to be sorted out later, teams that work with the system daily but had no part in building it, and documentation that describes how the thing was made rather than how to keep it running.

These delivery-scoped partnerships are still common for technology projects, and although they make solving Day 2 problems particularly difficult, for some kinds of software the gap they leave can be manageable. For AI, it tends not to be. An AI system in production has an additional set of ongoing requirements to conventional software, and these requirements don't always announce themselves when things start to go wrong. Most organisations who build AI with a scope that ends at delivery don't fully reckon with what that means until they're already in it: the partner has moved on, and the system they're left running is harder to maintain, govern, and keep reliable than they'd anticipated.

This is the delivery gap: the space between a system that has launched and an AI capability that runs reliably and within the rules. The organisations that close that gap tend to share a common approach: they work with partners who were thinking about month twelve from day one.

In 2026, closing the delivery gap is more urgent than ever. From 2 August 2026, the EU AI Act's foundational transparency rules take effect. While major high-risk obligations are deferred to late 2027, the runway is short. 'Launched' is no longer the finish line, and for organisations whose AI partnerships were built around a delivery endpoint, that's a real gap to close.

The provisional agreement also introduces a fixed timeline for the delayed application of high-risk rules: the new application dates would be 2 December 2027 for stand-alone high-risk AI systems and 2 August 2028 for high-risk AI systems embedded in products.

Furthermore, the provisional agreement reinstates the obligation for providers to register AI systems in the EU database for high-risk systems, where they consider their systems to be exempted from classification as high-risk. It also reinstates the standard of strict necessity for the processing of special categories of personal data for the purpose of ensuring bias detection and correction.

The provisional agreement postpones the deadline for the establishment of AI regulatory sandboxes by competent authorities at national level until 2 August 2027 and reduces the grace period for providers to implement transparency solutions for artificially generated content from 6 months to 3 months, with the new deadline set on 2 December 2026. The deal between the co-legislators also clarifies the competences of the AI Office for the supervision of AI systems based on general-purpose AI models where the model and that system are developed by the same provider by listing exceptions where national authorities remain competent, including law enforcement, border management, judicial authorities and financial institutions.

Legal and industry analysts stress that the delay is not a “snooze button” for compliance. The core requirements – such as data governance, technical documentation, and human oversight – remain unchanged. The primary recommendation from the EU AI Office and legal firms is to treat this 16-month extension as a “build period” rather than a wait period, as the complexity of achieving compliance (especially for high-risk systems) often requires significant operational overhauls that take over a year to implement correctly.

Strategy, engineering, operations: why AI partners must cover them all

The delivery gap appears so regularly because most AI partnerships are scoped around one phase of a three-phase journey. These partial scopes lead to three distinct failure modes in the long-term success of AI projects.

The first failure mode is when the consultancy stops at strategy. They produce a coherent roadmap, a well-argued business case, and a genuinely compelling vision of what the product could do. But then the engineering reality becomes somebody else's problem. The questions that determine whether an AI system actually works in production, like how the model behaves on real data rather than clean test sets, what inference costs look like at scale, and how it connects to the systems the business already runs on, need someone in the room who understands them from the start. Gaps between strategy and engineering reality can cause problems in any technology project, but in AI they often surface later and cost more to fix.

The second is the delivery partner that stops at launch. The system launches, it works on day one, and the engagement closes. But a live AI system is not a finished product. Unlike conventional software, it can degrade in ways that aren't immediately visible, including model drift, guardrails that require active maintenance as usage patterns change, and outputs that look plausible but are no longer sound. Keeping it running well requires ongoing monitoring, active governance, and someone who remains accountable for whether the decisions it's making are still the right ones. But if none of that was in scope, the business is handed something that is running and left to work out how to keep it that way.

The third failure mode is harder to see in advance, but often a significant issue in practice: making modern AI work smoothly with the legacy systems that the business is built on. Research in The Digital Sovereignty Trilemma found that 41% of IT leaders say their inability to move legacy business applications is actively holding them back. This is a real, widespread challenge. Even organisations not actively restrained by legacy systems are unlikely to have greenfield infrastructure. Most will already have decades of systems, processes, and data architectures that weren't designed with AI in mind.

Legacy integration is also where delivery partnerships most commonly leave unfinished business. The hardest connections – between modern AI capability and the older systems that run other functions – tend to get solved just well enough to launch, with the deeper work deferred. Once the engagement closes, that deferred work becomes the organisation's to manage.

End-to-end capability should include the same partner owning all these stages. Real operational intelligence means strategy design that's grounded in engineering reality, engineering that's designed to work on day one, and an operational strategy that's built in from the outset to keep systems running, compliant and improving over time. And, critically, it also means a partner who helps identify and design use cases that will create genuine business value, not just ones that are technically feasible.

Building lasting AI capability at a world-leading education provider

What closing the delivery gap actually looks like in practice goes beyond getting the technology right. We were approached by a world-leading education provider who operate in nearly 200 countries and provide digital content, assessments, and technology-powered learning solutions at global scale. This organisation's engineering division came to Insight AI asking a question most organisations don't think about early enough: not just whether AI would work, but whether their own teams would be equipped to run it safely, govern it properly, and keep improving it themselves.

When Insight AI began working with this organisation's engineering teams, an initial survey of the engineers taking part revealed something that headline AI adoption figures tend to hide. Most of the team were already using AI tools regularly, so the numbers looked healthy. But this is a pattern seen across many organisations: regular use of AI tools doesn't necessarily mean effective or safe use. Digging into how the tools were actually used revealed that structured evaluation of AI outputs was inconsistent, and the more sophisticated techniques that tend to generate the most meaningful gains were underused.

Even more significant was what was holding people back. Security and compliance anxiety was the single most cited barrier, named by engineers across every role and function. People who were uncertain about what data was safe to share with AI were either avoiding certain tools entirely or making their own judgement calls with no consistent framework to guide them. The problem wasn't carelessness; it was the absence of the standards, tooling guidance, and confidence needed to work safely. Deploying more technology wouldn't fix it.

What changed

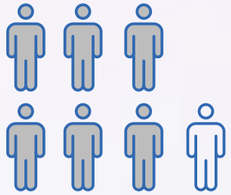
The programme set out to build what the teams had been missing: the infrastructure for safe, effective AI use. This happened through a combination of workshops grounded in real engineering and QA work, embedded coaching in live codebases and test suites, and a set of governance artefacts – usage guides, decision frameworks, prompt libraries, best-practice playbooks – designed explicitly to outlast the programme. Rather than leaving those standards to fade once the engagement ended, an internal champion network was built, with the most experienced practitioners taking visible ownership of the emerging standards and a teaching role in the coaching phase.

The governance figure is the one that matters most in the context of the 2nd August deadline. By the end of the programme, senior staff were naming specific controls: confirmed enterprise licence configurations, anonymised snippets for sensitive code, PR disclosure as standard practice. One team caught a write-capable tool overwriting ticket content before it became a live incident – the kind of practical risk awareness that doesn't come from a policy document, and takes months to build.

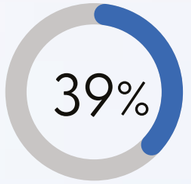
"Read every line. AI generates plausible-looking code that can test the wrong thing entirely. If you can't make the test fail with a bad input, it's not testing anything."

QA Engineer

The results went well beyond what had been targeted at the outset:



6 out of 7 developers now say AI tools save them real time in the actual work they deliver



39% reduction in lead time to first code commit, compared to a 15-20% target

2-8 hours saved per automated test created

10-30 minutes saved per bug report

From adoption to agency

The strongest evidence that this capability genuinely transferred was what happened after the programme ended. The QA team, the group that had rated themselves most conservatively at the start, went on to build two agents targeting specific problems in their own delivery process: one that reduced manual accessibility compliance work from up to 38 hours per page to around 30 minutes, another that cut design review time per component from up to three hours to under ten minutes. Neither was built by InsightAI. They were built by the team's own engineers, using capability that now belongs to that organisation.

Looking to build this kind of internal AI capability in your own teams?

Our specialists can walk you through this programme's approach – how capability transfer, governance, and real-world application were structured to last beyond delivery.

[Speak to an AI capability specialist]

Engineering with the end in mind

This programme demonstrates what it looks like to build lasting capability at the organisational level – equipping the people who run an AI system to use it safely and govern it well. The same question applies at the engineering level: whether the system itself was built to be operated, adapted, and governed over time, or simply built to launch.

We put those principles into practice ourselves when we built AURA, an AI-powered knowledge platform designed to solve a real internal problem: turning completed project work into client-facing case studies was slow, manual, and often meant the content wasn't ready when a sales opportunity needed it. With AURA, our sales teams have access to relevant proof points when they're working on a deal, and can use it to create narratives in multiple languages that show clients our expertise, leading to direct commercial gain. But building it also gave us the opportunity to apply the same engineering principles to our own systems that we'd apply to a client engagement.

The principle behind those decisions is straightforward: the choices that determine whether a system can be trusted over time are rarely visible in a demo. They show up months later, when the data the system handles becomes more sensitive, when usage scales, when the regulatory landscape tightens. So rather than treating governance and compliance as things to be added once the system was live, we built them in from the start. We made sure that data is anonymised by default, and access is governed by role. The infrastructure changes are auditable and reproducible rather than manually applied and hard to trace. And operational monitoring is built in, not bolted on.

None of these choices change what the system does on day one, but they determine whether AURA can be operated responsibly at scale, adapted as requirements change, and governed under increasingly demanding conditions – including the obligations the EU AI Act now makes ongoing rather than one-off.

AURA is still running, has kept evolving, and is governed by the same foundations we put in place at the start. The platform has now developed to create broader themed or industry-focused narratives from groups of relevant case studies – something that wouldn't have been possible to build on top of a system that hadn't been architected to grow.

What compliance actually requires

Even with high-risk deadlines deferred to December 2027, the foundations cannot be built overnight. Compliance requires continuous documentation, monitoring, and human oversight—not a one-time sign-off. They require ongoing conformity: continuous documentation, monitoring, human oversight of automated decisions, and clear accountability structures that can be demonstrated rather than just described. For organisations whose AI partnerships end at delivery, meeting that standard is an operational challenge they won't be set up for.

What the education provider's AI capability programme illustrates is what regulatory readiness looks like in practice: the 90% of senior staff who can name specific controls and apply them in context, the squad that caught a governance risk before it became a production incident, the standards, governance materials, and champion network that continue to govern the organisation's AI use as the technology keeps evolving.

That kind of operational foundation is also, increasingly, a commercial advantage. **The Digital Sovereignty Trilemma** found that 55% of organisations name regulatory complexity – GDPR, DORA, the AI Act – as one of their greatest strategic challenges. Where most find something hard, the ones who handle it well gain an advantage: 60% of clients now demand compliance proof as part of procurement, and 43% of organisations have used sovereignty credentials to win or retain business. The organisations that can demonstrate ongoing compliance, not just point to a delivery sign-off, are the ones that will hold an advantage as obligations continue to grow.

The full distance

Our education provider's AI capability programme and our Aura project both show what it looks like to build for the long term in practice, but arrived at from different directions. The AI capability programme built internal capability – the standards, governance, and confidence to keep improving – so that when Insight AI left, the organisation could carry it forward. Aura was built to last from the start, with the decisions that determine long-term operability made during the build rather than deferred to later. Both came down to the same thing: who is thinking beyond delivery before delivery begins?

For most organisations, that question gets answered long before the system goes live. It depends on whether the partner who built it was thinking about operations from the start: whether the strategy accounted for engineering reality, whether the engineering was designed for what would need to be operated, and whether ongoing responsibility was part of the engagement rather than a conversation for later.

That continuity of ownership – across strategy, engineering, and the ongoing work of keeping a system running – is what end-to-end actually means.

Want to see how these governance and engineering principles translate into a live system?

We can walk you through how AURA was designed for long-term operability including compliance by design, monitoring, and scalable architecture choices.
[Request a walkthrough of AURA]

Read Still Piloting? Here's How to Get AI into Production Faster for the other side of this challenge – moving AI from pilot to production, and what it takes to get there.

se.insight.com/ai