

AI DEPLOYMENT INSIGHTS

Why Conversational AI Wins Pilots and Loses in Production

A briefing for senior leaders transforming high-volume customer service operations.



THE ARGUMENT IN 60 SECONDS

Most conversational AI pilots succeed. Most production deployments don't deliver what the pilot promised. The gap isn't bad technology - it's structural. Pilots are designed around conditions that don't exist at scale: curated data, limited scope, dedicated teams, and metrics defined after seeing early results. Production exposes everything the pilot was shielded from.

The organisations that scale AI successfully don't run smoother pilots. They run pilots that surface problems early - while the team is focused, the stakeholders are engaged, and the cost of fixing things is low. A pilot that finds nothing wrong has almost certainly looked in the wrong places.

THE MOMENT WE ARE IN

AI in customer service is no longer a strategic option - it's an operational reality.

Microsoft's 2025 Work Trend Index found that:

46%

of leaders say their companies are already using AI agents to fully automate workflows

81%

expect agents to be moderately or extensively integrated into their company's AI strategy within 12-18 months

71%

of leaders at the most AI-mature companies say their company is thriving, compared with just 39% of workers globally.

The gap between leaders and laggards is widening fast, and the question for most organizations isn't whether to deploy AI in customer service anymore, but how to make it stick.



WHY PILOTS SUCCEED

Pilots are structurally designed to win. Four conditions make them look good - and none of them survive contact with production.

Limited scope

Pilots target best-case use cases: high volume, well-documented, clean flows. The messy, ambiguous, emotionally charged interactions never make it into the test.

Clean data

Curated inputs. No edge cases, no legacy inconsistencies, no contradictions between what the knowledge base says and what the live process actually does.

Controlled environment

A dedicated team with full attention and no competing priorities. Production means shared resources, organisational politics, and a dozen other things demanding the same people's time

Favourable measurement

Success metrics often get defined after seeing early results, not before. The pilot is measured against a bar it helped set.

These conditions don't just make pilots look good - they prevent pilots from doing their actual job. A pilot exists to learn. It's the one window where you have dedicated people, senior attention, and the budget to course-correct. Every problem found during a pilot is a problem you don't discover in production, when the team has dispersed, attention has moved on, and fixing things costs an order of magnitude more. The discomfort of finding problems early is the entire point.

THE PATTERN EVERY LEADER RECOGNISES

The pilot looks strong.

Automation rates improve, service levels hold, the business case is clear, stakeholders are excited, and the board approves the next phase.

Then production arrives.

Escalation rates rise. Complexity increases. Operations struggle to keep pace. Costs creep back up. The AI project gets quietly shelved.

This isn't an exception.

It's the default trajectory when pilots are designed to validate a decision already made, rather than to stress-test whether that decision holds under real conditions.



SIX THINGS THAT BREAK BETWEEN PILOT AND PRODUCTION

Each of the following draws on real deployments Automaise has led across insurance, ecommerce, FMCG, energy, and financial services. The pattern repeats regardless of sector. In every case, the issue could have been caught during the pilot - and in one case, it was. The difference is whether the pilot was designed to find these problems or to avoid them.

1. Pilot design

What happens

The pilot runs on structured hours, scoped volumes, cooperative customers. It captures the easy conversations and avoids the hard ones - peak load windows, high emotional intensity, edge-case scenarios.

In practice

An insurance deployment ran a pilot with defined operating hours and controlled scope. Metrics looked strong. But the pilot never tested against peak volume or the emotional range of real claims and coverage calls. When production brought real load and real feeling, the experience didn't match. The pilot had been too comfortable to be predictive.

The fix

Design the pilot to be uncomfortable. Test peak hours, real customers, the full emotional range. Align success criteria identically between pilot and production. A pilot that reveals cracks in the experience is doing exactly what it should - you have the team assembled, the attention from leadership, and the mandate to fix things. That window closes the moment the pilot is declared a success, and everyone moves on. If the pilot doesn't include a graduated rollout plan and a decision framework for expansion, even a genuinely successful pilot still stalls.

If you skip this

The pilot succeeds on paper. Rolled out to realistic conditions, results collapse.



2 • Knowledge base quality

What happens

The pilot runs on curated content. Production means the full knowledge base - outdated articles, internal contradictions, content structured for human agents rather than AI retrieval.

In practice

An e-commerce retailer had multiple KB articles on returns policies that contradicted each other - and contradicted the logic baked into the conversational agent's flow. Separately, escalation articles explained how to reach a human agent, but the agent had been updated to block that path. Customers read one thing and experienced another. The root cause wasn't hallucination. It was that the KB had never been audited against the live agent logic. Content and conversation design lived in separate worlds, and nobody was maintaining the alignment.

The fix

Audit content against live processes before go-live. Treat knowledge base alignment as ongoing hygiene, not a one-off task. A well-designed pilot would deliberately stress-test the full KB - not a curated subset - and treat every contradiction it surfaces as a cheap fix. Finding this during the pilot, with the project team focused and the content owners engaged, is when you can actually resolve it. Finding it in production means hundreds of wrong answers served before anyone raises the flag.

If you skip this

Confident, wrong answers at scale. The AI didn't hallucinate - it was fed contradictions and asked to sound confident about them.



3. Systems integration and data readiness

What happens

The pilot runs in a simulated or contained environment. The experience feels fast and accurate because real systems aren't connected yet.

In practice

A FMCG company did this right. They integrated with their Salesforce sandbox during the pilot - not after. The team discovered that the volume and nature of GET requests made for a slow, clunky conversation flow, unviable for real-time customer interaction. They caught it, optimised the queries, and fixed it before production. The pilot surfaced the problem, which is exactly what a pilot should do.

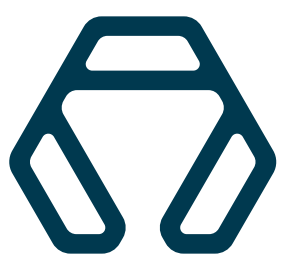
Most pilots take the opposite path: mock integrations entirely or skip them. The conversation feels snappy because nothing real is connected. Production is the first time anyone tests the real plumbing.

The fix

Integrate with realistic systems during the pilot. Bottlenecks found early are a success, not a failure. This is the clearest example of the principle: the pilot exists to find problems while you still have all hands-on deck to fix them. A pilot that surfaces an integration bottleneck and resolves it has done more for production readiness than a pilot that reports clean metrics because nothing real was connected.

If you skip this

Production conversations degrade - slower, unreliable, inconsistent. The AI looks broken when the problem is what sits behind it



4 • Use-case fitness and assumed portability

What happens

AI is mandated centrally, but operations vary across markets, segments, and business units. What works in one context doesn't transfer when the underlying dynamics differ.

In practice

A FMCG organisation mandated AI centrally and built a POC for B2B use cases - order taking and invoice requests - scoped against one European market with decentralised ordering and close buyer-supplier relationships. That model didn't transfer to the UK, where buyers are major distributors and supermarket chains with minimal direct supplier contact. The use case that made sense in one market had no operational basis in another, but the central mandate assumed it would.

A similar pattern appeared in an energy group: a group-level AI initiative where operational realities and process maturity varied across business units. What worked in one part of the group hit friction in another.

The fix

Validate fit locally before assuming portability. Central strategy sets direction; local discovery confirms the fit. If the pilot phase includes even a lightweight validation against the next target market or business unit - testing the assumptions, not just the technology - the mismatch surfaces early enough to adapt the approach rather than force a rollout that was never going to land.

If you skip this

The AI works technically but not operationally. Local teams resist, adoption stalls, and the initiative loses credibility.



5. Compliance, regulation, and infrastructure

What happens

The pilot operates below the compliance radar - limited data, limited users, local approvals. Production triggers the full regulatory and security machinery. Both sides defer these conversations to keep momentum.

In practice

A financial services deployment for a European banking group ran a contained pilot - a few hundred clients, local jurisdiction, local approvals. Enough to prove the concept. The moment production-scale questions arose - client data handling, systems integrations, security posture - the full compliance and infosec machinery activated. The deployment couldn't run on the vendor's cloud. It needed to conform to the bank's specific security perimeter, unique to that institution and different again at the next bank.

Both sides knew these conversations existed. Both sides deferred them. The vendor wanted ink on paper. The client's business team wanted momentum. Nobody wanted to invite infosec to the table early because it would slow everything down. The cost didn't disappear - it moved to where it was ten times more expensive.

The fix

Have the hard conversation early, even when commercially inconvenient - on both sides. The pilot is the right moment to involve compliance and infosec, precisely because you still have the project team, the executive sponsor, and the budget to absorb the impact. Yes, it slows the pilot down. But a slower pilot that arrives at production ready to deploy is worth more than a fast pilot that hits a wall the moment it tries to scale.

If you skip this

Production hits a wall. Architecture needs redesigning. Hundreds of hours of unplanned compliance and infrastructure work. Timeline and budget blow out.



6. Operational ownership and maintenance

What happens

The pilot is run by senior people on both sides. Production requires structured handover with governance, RACI, and training across operational, managerial, and strategic levels.

In practice

An e-commerce retailer had no sustained ownership post-deployment. Degradation accumulated fast. The vendor was called back in to clean up - the classic scenario where chaos was visible almost immediately because nobody was watching.

In an energy utilities deployment, the ambiguity was structural from the start. The RFP didn't scope ongoing use-case build continuity. Multiple teams - internal, outsourcers, consultants - were in play with no clarity on who inherits what when the vendor contract expires. People were being trained, but nobody had defined who carries the torch.

The fix

Treat change management as a core deliverable. Define RACI before the pilot ends. Train the right teams while knowledge is fresh. The pilot is when senior people from both sides are engaged and the operational model is being shaped in real time. That's the moment to define ownership, build the governance structure, and transfer knowledge - not after the project team has moved on and the institutional memory has already started to fade.

If you skip this

Without clear ownership, chaos from day one. Monitoring, maintenance, and issue resolution break down immediately.



CASE STUDY • FINANCIAL SERVICES

Real-Time Intent Routing Replaces Legacy IVR

A leading European bancassurance provider transforms inbound contact center routing with real-time AI intent classification.

Real-time
routing

Zero IVR
menus

90%+
Intent classification
accuracy

The challenge

Traditional IVR forced customers into generic options, causing transfers, cost, and frustration. Manual call distribution was inflating handling times across the board. And the deployment had to meet enterprise-grade security requirements aligned with DORA regulations.

The solution

Real-time intent classification replaced the IVR entirely - routing every call to the right team without menus or transfers. Talkdesk CTI and Salesforce CRM integration automated call distribution and case creation. The platform was hosted on Microsoft Azure private cloud, delivering security and governance by design.

The results

>90%

Intent classification accuracy in
call routing

Fewer Transfers

Reduced call transfers, improved
first-contact efficiency

Faster Resolution

Higher customer satisfaction and
stronger agent productivity



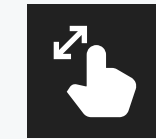
WHAT ORGANISATIONS THAT SCALE AI IN PRODUCTION DO DIFFERENTLY

Most organisations try to automate everything at once to justify the investment. That's exactly why they fail. The ones that scale start with one use case, prove it fast, and expand from confidence - not pressure. They also treat the pilot as a learning instrument, not a validation exercise. The goal isn't a clean scorecard - it's a list of everything that needs to be true before production, tested and resolved while the team and the budget are still there.



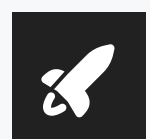
Choose the right entry use case.

The one with the highest volume, clearest scope, and fastest measurable impact. Start there. Prove it. Scale it.



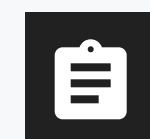
Start small but with expansion in mind.

The first use case is never the last one. Define and document where you want to go after the first launch.



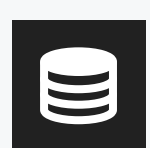
Measure impact from day one.

Define metrics before go-live. If you can't measure it within 30 days, you chose the wrong use case.



Assign clear operational ownership.

Named owners for optimisation, knowledge, and escalation. Not the vendor, not IT alone. That's what keeps production healthy after the project team moves on.



Prepare the ground: data, systems, and change management.

AI fails without clean data, integrated systems, and people ready to work alongside it. Map systems, processes, and people before kick-off. Define integrations before you build. Invest in change management from day one.



QUESTIONS TO TAKE BACK TO YOUR ORGANISATION

On pilot design

How do we design pilots that genuinely predict production performance, not just prove a concept?

What metrics should be defined from the outset to reflect real operational impact?

On operational ownership

Who owns AI-driven customer service operations once they are live?

What does a realistic and sustainable operating model look like in terms of roles, accountability, and ongoing optimisation?

On use-case prioritisation

How do we balance automation targets with customer experience - and where is the line?

How do we prioritise the right use cases from the start and ensure the right expertise is in place?



ABOUT AUTOMAISE

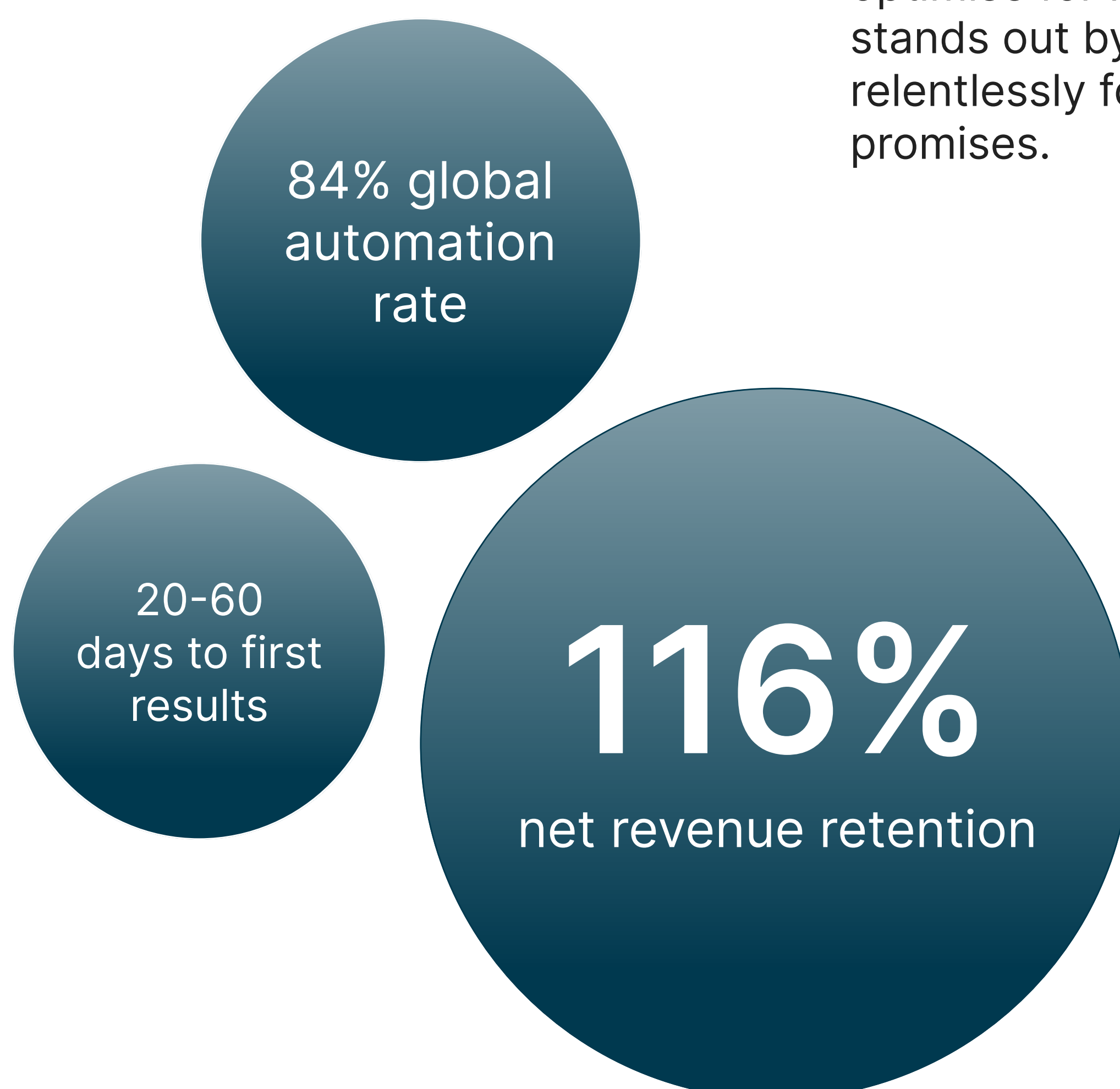
Our AI agents
don't stop
at answers.

They own
the resolution.

We help enterprises reduce the cost of customer service and deliver better experiences through AI agents that automate complex interactions across voice, email, chat, WhatsApp, and SMS.

Designed for high-volume, multi-channel environments, Automaise integrates with existing CRM and contact centre systems to automate end-to-end processes – not just simple interactions – delivering measurable impact in cost per interaction, average handling time, and customer satisfaction.

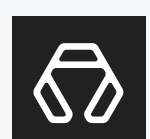
What sets us apart is our focus on value over hype: we don't just provide tools, we partner with our clients to identify high-impact use cases, deploy quickly, and continuously optimise for results. In a crowded AI market, Automaise stands out by being pragmatic, outcomedriven, and relentlessly focused on delivering tangible ROI rather than promises.



TRUSTED BY



GET IN TOUCH



Julian Degg

Sales Manager UKI & US, Automaise

julian.degg@automaise.com

+44 7500 056 844



Andrew Talas

Head of Solutions, Automaise

andrew.talas@automaise.com

+351 916 876 554