A Handbook for Public Service Leaders

AI Agents Handbook

Redesigning Public Services for the Age of AI Agents

Chris Pannell UKI VP Agentic Automation & Communications



How This Handbook Works

Contents

This handbook is not a guide to AI features.
It is a guide to redesigning public services in the age of AI agents.

Read the opening chapters to understand the shift. Use the diagnostic and capability sections to identify where transformation is possible. Then use the governance, measurement and scaling chapters to turn that opportunity into a repeatable operating model.

ORIENTATION	*Understand the shift*	*Page*
01	Foreword **The Pattern We Can't Ignore** Why most AI programmes fail to scale	03
02	State of the Nation **AI in UK Public Services** Where UK public services are today – and where they are not yet	04
03	The Strategic Fork **Two Paths Diverge** The difference between augmentation and true transformation	07
APPLICATION	*Find the opportunity*	
04	The Diagnostic **Finding Your Transformational Opportunity** How to identify the service opportunity worth pursuing	10
05	The Capability Toolkit **What You Need to Build a Real Service Not a Demo** What a real agentic service needs in order to work	19
06	Governance that Enables Scale **Designing Control that Accelerates Delivery** How to create control without blocking delivery	24
07	Measuring What Matters **Proving Service Transformation, Not Just AI Activity** How to prove the service has changed	28
08	Scaling Without Losing Control **Scale Patterns, not One-Offs** How to turn success into a repeatable operating model	32
CONCLUSION	*Turn it into an operating model*	
09	**Conclusion** What public services must protect and amplify	36

Foreword

The Pattern We Can't Ignore

One year on from the first edition of the AI Agents Handbook, the pace of innovation has only accelerated.

Across UK Government, there are encouraging signs of progress: live deployments, measurable impact, and teams proving that agentic technology can deliver real value.

And yet, something uncomfortable remains true. MIT-linked 2025 research suggests that only around 5% of custom generative AI pilots make it into production. In other words, there is a "5% club" – and most organisations are still outside it. In government, the pattern is disturbingly familiar. Pilots flourish in corners. Innovation may win awards, but enterprise-wide transformation remains elusive. The problem is not bad data, capability gaps or immature technology. It is something older, more familiar and more systemic: ***feature thinking.***

Feature thinking puts technology ahead of outcomes. It says: we have turned the AI on now where can we use it? Add a chatbot. Redesign a form. Embed a co-pilot. It is solution first thinking, and it is why so many programmes fail to scale.

Product thinking starts somewhere else entirely. It asks: what outcome are we accountable for? Where is friction creating failure demand? What repeat work could disappear? What would materially improve the lived experience of the human providing or receiving the service?

This is hard in government because ownership is fragmented. Policy owns legislation. Operations own delivery. Digital owns systems. Suppliers own contracts. Rarely does one person own the end-to-end service outcome. That is why agentic programmes stall when treated as IT-led automation initiatives. The strongest deployments are owned by the people accountable for the job to be done – **with technology enabling the outcome, not defining it.**

The organisations that make it into the 5% do not treat AI as another feature to deploy. They treat it as part of the operating model they now need to redesign.

This handbook is a blueprint for avoiding that trap – and for moving from pilot theatre to operational transformation.



Before embarking on any AI journey, one question matters above all:

Does this improve the lived experience of the human providing or receiving the service?

If you cannot answer the human question clearly, you are in the feature trap.

02 | State of the Nation

AI in UK Public Services



“The UK has made progress. But progress is not redesign.”

State of the Nation

AI in UK Public Services

The UK has moved beyond the question of whether artificial intelligence belongs in public services.

The AI Opportunities Action Plan is now in delivery. UKRI has committed more than £1.6 billion to AI research and innovation. Councils report that AI is now widely used or being explored. AI-assisted diagnostics are scaling in the NHS. Generative tools are embedded across departments.

Adoption is no longer the issue. Maturity is.

THE SHIFT IS REAL

Across central and local government, AI has moved beyond theory. There are funded programmes, live deployments and a growing body of experimentation. In some services, agents are already being used in production. In others, generative tools are supporting staff with drafting, summarising, search and case preparation.

This is no longer a fringe conversation.

What is striking, however, is that citizens are often moving faster than institutions. Government research published in early 2026 found that 73% of the public had used AI at least once in the past month, with usage highest among younger adults. The constraint is no longer public familiarity with the technology. It is the ability of institutions to operationalise it safely, usefully and at scale.

THE PATTERN REMAINS IMMATURE

Most deployments still sit at the augmentation layer. They improve knowledge work: summarising documents, drafting correspondence, analysing consultations, capturing meeting notes and surfacing policy guidance. These tools reduce friction and administrative burden. Reported gains are typically measured in minutes saved, task acceleration or staff efficiency.

Valuable, certainly. Transformational, not yet.

Research suggests that around 41% of public-sector time could theoretically be supported by generative AI. Yet the UK continues to face a substantial productivity gap linked to weak digital foundations and fragmented operating models. Public trust remains fragile too, with polling showing that lack of trust in AI remains one of the biggest barriers to adoption.

In short, experimentation is widespread. Operational redesign is still limited.



£1.6bn
UKRI commitment to AI research
UKRI · FEB 2026

95%
English councils using or exploring AI
LGA · 2025

73%
UK public used AI in the past month
DSIT / DCMS · JAN 2026

State of the Nation

Early Signals of the Next Phase

There are early signals of what more mature adoption looks like.

Several UK police forces are now deploying live digital agents, including Bobbi. Unlike summarisation tools, these systems are public-facing and workflow-integrated. They enable citizens to report incidents digitally, access information outside office hours and navigate services without defaulting to non-emergency call lines. Early deployments indicate reduced routine demand, improved quality of information capture and faster routing to the appropriate teams.

Internally, agentic systems are also beginning to emerge within collaboration platforms and service portals. These deployments connect AI reasoning to structured data and operational workflows, enabling task execution and completion rather than passive assistance. The difference is subtle but significant: AI moves from drafting to doing.

Public bodies have only two enduring assets with which to improve services: **their people and their institutional knowledge.**

AI should make that knowledge more accessible – to the public, to reduce avoidable burden, and to staff, so that low-value work can disappear and human effort can be redirected to judgement, empathy and complex casework.

The UK has made progress.
But progress is not redesign.

The next question is not whether AI can help. It is whether institutions are prepared to change how work happens.

Early Signal

From drafting to doing

- Public-facing, workflow-integrated agents
- Complex work returned to human judgement
- Connected to structured data and operational systems
- Routine demand deflected; information capture improves

The value grows when AI moves beyond helping individuals and starts changing how demand is handled.

03 | The Strategic Fork

Two Paths Diverge



“This is not a question of whether AI helps.
It is a question of where in the operating model AI is deployed.”

The Strategic Fork

Two Paths Diverge

This is the strategic fork.
Use AI at the edge of existing work, and the work gets faster. Use AI to redesign the service, and the operating model starts to change. Both approaches create value. Only one changes the system.

Path One

Augmentation at the Edge

AI makes existing work faster. The service remains fundamentally unchanged.

01 Productivity gains are real

02 Repeat contact remains largely unaffected

03 Cost per transaction is unchanged

04 Channel mix does not materially shift

05 Human escalation remains necessary

06 Citizen accessibility does not expand

Typical Result

The work becomes faster. The system stays the same.

Path Two

Orchestrated Agentic Redesign

AI is designed into the service itself. Demand is resolved, not simply processed.

01 High-volume interactions are handled end-to-end

02 Channel mix shifts toward digital solution

03 Repeat contact declines

04 Cost per interaction drops materially

05 Work is removed, not merely accelerated

06 Human judgement is reserved for exceptions

Typical Result

Low-value work is removed. The service model improves.

Why live systems matter

The strongest deployments are tested against real demand, real users and real operational pressure. Salesforce has done this internally with Agentforce: on Salesforce Help, Agentforce now resolves more than 68% of conversations.

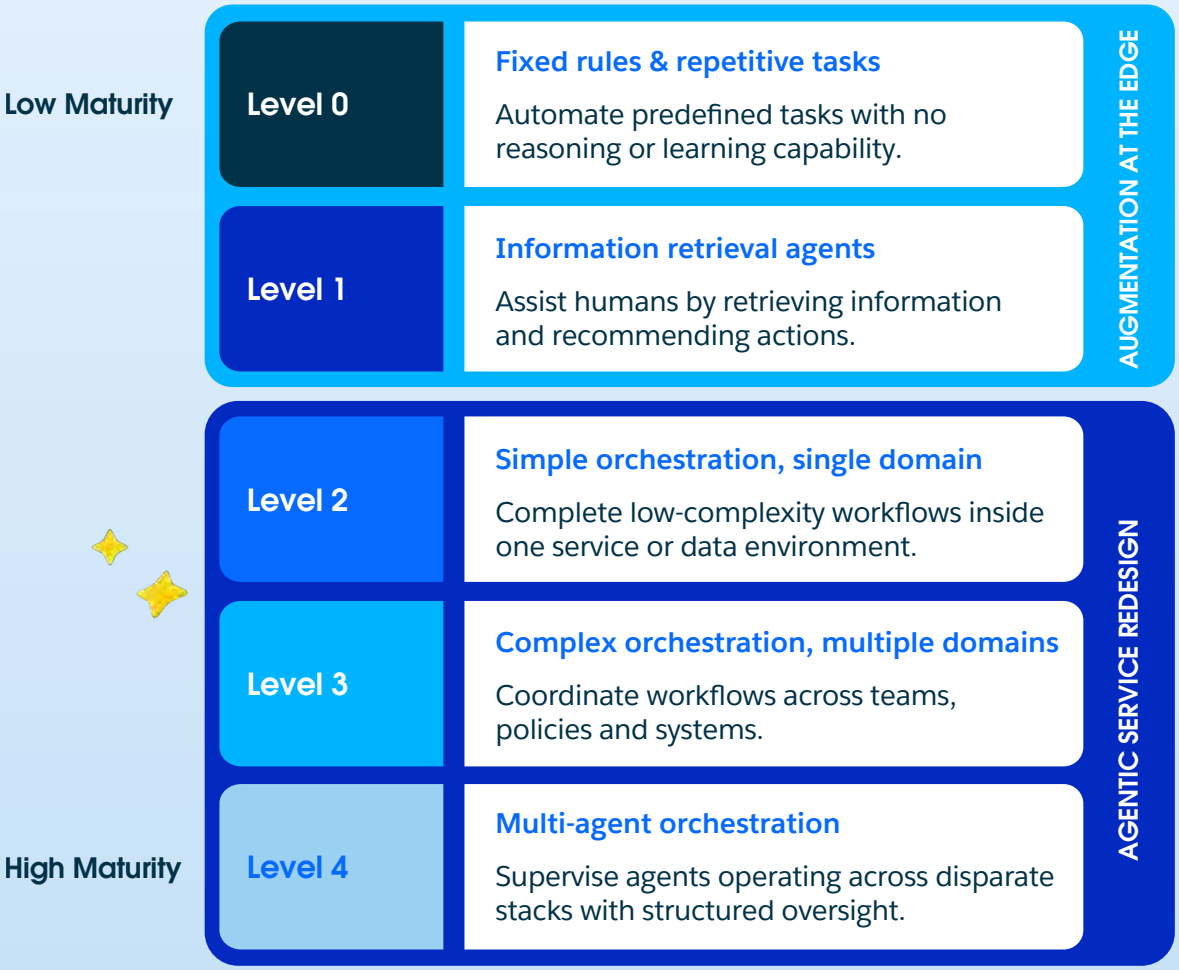
The lesson is simple: transformation happens in live systems, not pilots.

The Strategic Fork

The Maturity Shift

The difference between these two paths maps directly to organisational maturity.

At lower maturity levels, AI performs fixed or retrieval-based tasks. It recommends, drafts and summarises. At higher maturity levels, AI orchestrates across domains. It coordinates workflows, supervises other agents and operates across disparate systems with structured oversight.



The Inflection Point

Incremental gains are useful, but so often they cannot deliver returns beyond self-reported gains. But incremental gains do not compound into transformation. Transformation requires targeting the services generating the majority of demand and redesigning them around agentic workflows.

- **You cannot optimise your way to a tenfold reduction in cost-to-serve.**
- **You cannot summarise your way to 24/7 multilingual service access.**
- **You cannot draft your way out of structural backlogs.**

That is where the next phase begins.

04 | The Diagnostic

Finding Your Transformational Opportunity



“The economic value comes from removing avoidable demand, improving access and redeploying human judgement where it matters most. That is the distinction the diagnostic is designed to surface.”

The Diagnostic

Finding Your Transformational Opportunity

You have chosen a service-led transformation path, not a feature-led one. The task now is not to find a use for AI. It is to identify where redesign will most materially improve the service. Not with the easiest problem. Not with the safest pilot. And not with whatever the technology happens to make possible. Start where the opportunity is greatest: where demand is highest, friction is deepest, and the economics of the service are most broken.

That requires discipline. It requires looking at data you may not want to see. And it requires resisting the temptation to begin with technical capability rather than operational necessity.

This diagnostic is designed to help you do exactly that.

Work through it with service owners, operational leaders, frontline teams, digital leads and data owners. The purpose is not to prove that AI is interesting. It is to identify whether an agentic redesign would materially improve the service.



Six Steps With Output

01 **Define the service**
...in citizen terms.

02 **Surface friction & failure demand**
Where demand is generated.

03 **Quantify the baseline**
Agreed economics and performance.

04 **Define the target state**
Material, measurable shifts.

05 **Apply the agent filter**
A deliberate decision gate.

06 **Be brave about complexity**
Be explicit about trade-offs.

01 Define the service in citizen terms

Before you build an agent, write down the service you believe you deliver. Then rewrite it from the perspective of the person receiving it – or the frontline team delivering it.

Worked Contrast

X **"We process Blue Badge applications."**

"We enable disabled citizens to access the parking permissions they are entitled to, so they can maintain independence."



Services designed around internal requirements optimise internal requirements: speed, compliance, security, cost and closure. Services designed around human outcomes can be redesigned around real need. **If you cannot articulate the human outcome clearly, stop. You are not ready to build.**



Output of this Step

Service definition

- A one-sentence service definition written in citizen or frontline-outcome terms

Alignment check

Ask policy, operations and digital teams to define the service independently. If the answers differ, you have surfaced the first constraint: the organisation is not yet aligned on what the service exists to achieve.

02 Surface friction and failure demand

Now confront the data. Bring complaints logs, contact-centre analytics, abandonment rates, escalation data and staff feedback into the room.

Five Questions to Ask

- Why do citizens contact us repeatedly?
- What happens when the primary channel does not work?
- Where do processes fail, stall or loop back?
- Who is excluded by language, time or digital access?
- Where do people drop out?
- What work is repeated because the service failed first time?

Frontline stories reveal the workarounds. Data shows whether they are systemic.

In most services, a large proportion of demand clusters around a small number of recurring issues. Some are visible in the data. Others first appear through complaints, staff feedback or informal workarounds.

Many of these issues are informational: status updates, eligibility clarification, document requirements, “what happens next?”. Others are caused by poor communication, unclear policy interpretation or broken handoffs between teams.

This is not noise. It is the shape of the opportunity.



Output of this Step

Demand profile

- Top 10 call or contact drivers
- Top 3 repeat-contact causes
- Top 3 failure / abandonment points

03 Quantify the baseline

Transformation requires numbers. Define the current state in operational terms.

The Baseline You Must Capture

It may be necessary to baseline these metrics per channel or surface

01 **Cost per interaction**

02 **Cost per completed case**

03 **Average processing time**

04 **Repeat-contact rate**

05 **Escalation rate**

06 **Channel mix across phone, digital and in-person**

07 **Language coverage and service hours**

This is the baseline against which all future claims must be tested.

If “better” is not quantified, it cannot be measured.

If it cannot be measured, it cannot be improved.

And if it cannot be improved, it will remain a pilot story rather than a scalable service.

A weak baseline creates weak claims. A strong baseline creates the conditions for real proof.

Output of this Step

Baseline economics

- A baseline metrics table
- Agreed across service, ops and digital

04 Define what "better" means

Now define the future state. Not vague aspirations. Not "improved experience." Not "more efficient." Define measurable shifts.

Six Measurable Steps to Aim For

Processing Time
6 weeks > 48 hours

Cost Per Case
£120 > £15

Repeat-Contract Rate
40% > 5%

Channel Mix
80% phone > 30% phone

Service Availability
Office Hours > 24 / 7

Supported Languages
English only > 5 major languages

These measures should reflect something a human actually feels: – faster answers, fewer handoffs, clearer status, better access – **but they must be expressed in operational terms**, not soft sentiment or self-reported gains.

If the improvement does not materially change cost-to-serve, repeat demand, accessibility or backlog, **it is not transformational.**



Output of this Step

Target state

- Three to five target metrics that would represent a material shift in service economics, accessibility or operational performance

05 Apply the agent filter



Now ask the critical question: Would an agent materially move these metrics? If friction is caused by unclear legislation, an agent may simply automate confusion.

Criteria to Consider

PROCEED WITH CAUTION

01 **Policy unclear**

02 **Requires human judgement**

03 **Trust is broken**

STRONG FIT

01 **High-volume status queries**

02 **Repeated eligibility questions**

03 **Access constraints (hours, language, channel)**

04 **Structured workflows**

If delay is caused by a statutory approval step requiring human judgement, an agent cannot remove it. If trust in the process is broken, automation alone will not rebuild confidence.

But often the answer is yes.

- If 60% of calls are status updates, an agent can remove them.
- If applications fail because people misunderstand eligibility, an agent can guide them in real time.
- If access is restricted to office hours, an agent can extend availability immediately.
- If routine handoffs are slow and inconsistent, an agent can route and validate them deterministically

This is your decision gate.

Proceed only if the agent directly affects the high-volume, high-cost drivers you identified.

Output of this Step

Decision gate

- A clear Proceed / Do Not Proceed decision
- Rationale linked to the target metrics

06 Complexity and Trade-Offs



The biggest transformational opportunity will almost always be complex. Be brave about complexity, and explicit about trade-offs.

Trade-Offs to Make Explicit

Trade-Off	Example Decision
Speed vs. Accuracy	Instant decisions for standard cases; exceptions routed to human review
Automation vs. Oversight	60% of standard cases automated end-to-end; 40% remain human-led
Consistency vs. Discretion	Policy-as-code used for routine eligibility; discretion reserved for hardship or vulnerability cases
Investment vs. Savings	Integrate status and document systems now to remove repeat contact later

It will span systems. It will involve policy nuance. It will include edge cases. It may be the problem your organisation has postponed for years.

That is precisely why it matters.

But transformation is not just about tackling complexity. It is about choosing tradeoffs deliberately. Every meaningful redesign forces decisions. Make these choices explicit.

If a service drives 40% of your call volume, controlled automation may be rational. If it involves safeguarding, enforcement or high-risk decisions, the balance will look different.

Transformational organisations are not reckless. They are deliberate. They say: “We will automate the high-volume, rules-based cases. We will retain human judgement for the complex and sensitive ones. We accept short-term integration cost to permanently reduce avoidable demand.”

Output of this Step

Trade-off frame

- Documented list of explicit trade-offs
- Defined automation threshold
- Clear escalation and oversight model

Named accountable owner for risk and performance

Agreed risk tolerance aligned to service criticality

If you cannot articulate these clearly, you are not ready to deploy at scale.

The Diagnostic

Worked Example: Housing Benefit Applications

A local authority applies the diagnostic to a high-volume housing benefit service.

Diagnostic Lens	What the Team Found	What it Means
Service definition	Help low-income residents access financial support quickly and accurately so they can secure stable housing.	The service is framed around stability, speed and and clarity - not form processing.
Friction & failure	60% of inbound calls are status enquiries. 40% of applications are incomplete or incorrect. Support is office-hours only.	Demand is being created by uncertainty, poor guidance and limited access.
Baseline economics	£120 per completed case. 6–8 week processing time. 40% repeat contact. 80% of contact via phone.	A high-cost, high-volume service with clear demand concentration.
Target state	Standard cases processed within 72 hours. Repeat contact reduced below 15%. 24/7 multilingual status access. Cost per standard case reduced materially.	The ambition is a measurable shift in both economics and accessibility.
Agent opportunity	An agent could guide eligibility checks, validate documents, provide status updates and route complex cases.	High-volume, rules-based work can be absorbed digitally; human staff focus on exceptions.
Decision gate	Strong fit for an agentic redesign in the standard case journey. Poor fit for complex discretionary decisions.	Proceed with bounded automation and explicit human escalation.
Trade-offs	Up to 55% of straightforward cases automated. 45% remain human-led.	Automation is bounded, with clear oversight and escalation paths.
Projected shift	Fewer avoidable status calls. Faster processing. Lower cost-to-serve. More caseworker time for complex assessment.	The service model changes – not just the speed of the work.

Why This Is Transformational

The economic value comes from removing avoidable demand, improving access and redeploying human judgement where it matters most. **That is the distinction the diagnostic is designed to surface.**

05 | The Capability Toolkit

What You Need to Build a Real Service Not a Demo



“You do not get transformation by starting with the interface. You get it by building the capability stack underneath it.”

The Capability Toolkit

What You Need to Build a Real Service, Not a Demo

The next question is not which model to use. It is what capabilities you need in order to redesign the service safely, credibly and at scale. This is where many programmes drift back into feature thinking – choosing a model, building a demo and assuming progress. Transformation does not come from the model alone. It comes from the surrounding capability: the knowledge, workflows, controls and interfaces that allow an agent to do something useful in the real world.

The practical question is no longer theoretical.

What must be true for an agent to complete meaningful work inside this service?

The answer comes down to five capabilities that must exist together.

A real service is not built from a model alone.
It is built from a set of capabilities that enable knowledge, action, coordination and trust.



Key Insight

You don't get transformation by starting with the interface.

You get it by grounding the service in trusted knowledge, enabling action, designing it to work together, making it usable and ensuring it remains visible and governable once live.

The Capability Stack

05 **Human handoff & observability**
Know what happens, when to escalate and how to improve.

04 **A consistent interface for work**
Bring the right context, actions and next steps into one seamless experience.

03 **Composable design**
Build in parts, reuse what works and scale with confidence.

02 **Action, not just answer**
Take meaningful action, update records and move work forward.

01 **Trusted knowledge**
Know what happens, when to escalate and how to improve.

Foundation
Transformational Service Opportunity

01

CAPABILITY ONE

Trusted knowledge

An agent should not rely on generic model recall when a service depends on current policy, procedural nuance or local interpretation. It must be grounded in trusted organisational knowledge: guidance, decision criteria, service standards, process rules and, where appropriate, case history.

This is what turns a fluent model into a useful worker. If the agent cannot reliably access the right knowledge, it may sound impressive while remaining operationally weak. It may generate plausible responses. It will not generate trust. In practice, this often matters more than model choice.

Output of this Step

- A trusted knowledge map for the service
- A defined set of approved sources
- A clear owner for each knowledge source
- A process for updating, retiring and validating content

Common trap: A fluent answer is not the same as a trusted answer.

If the service depends on policy, eligibility or operational rules, the agent must be grounded in approved knowledge sources. If the answer cannot be traced, it cannot be trusted.

02

CAPABILITY TWO

Action, not just answer

Many AI deployments stop at response generation. They answer a question, suggest a next step or summarise a case. That is useful, but it is still augmentation.

Transformation begins when an agent can take action within a governed boundary. That might mean checking an application status, validating a document, updating a record, triggering a notification, routing a case, requesting approval or escalating to a human with full context.

This is the point at which AI moves from describing work to participating in it. A good test is simple: *if the agent can only talk about the process, but cannot engage in it, then the service has not yet been redesigned.*

Output of this Step

- A defined set of agent actions and permissions
- A clear boundary between what the agent can suggest, do and escalate
- A write-back model for how actions update systems of record

03

CAPABILITY THREE

Composable design

One of the most common mistakes in early agent design is to ask one agent to do everything. A better pattern is composability. Design discrete agents or capabilities for discrete jobs. One may handle knowledge retrieval. Another may validate a document. Another may manage triage. Another may orchestrate handoff and escalation.

This keeps the scope understandable, makes iteration faster and reduces the risk of brittle failure. A composable design also makes scaling easier. Instead of replicating a monolithic agent from service to service, you can reuse patterns, actions and workflows across multiple contexts.

Output of this Step

- A composable agent design or capability map
- A clear description of each agent's role
- Defined handoffs between agent capabilities
- Reusable components for future services

Common trap: Context windows are not architecture.

Adding more input text may help in narrow cases, but it does not replace clear roles, bounded actions and defined handoffs between capabilities. If one agent needs every rule, exception and workflow to work, the design is probably too broad.

04

CAPABILITY FOUR

A consistent interface for work

Even when the right knowledge and actions exist, the experience still fails if work remains scattered.

A good agentic experience brings signal, context and next action into one place. **The principle is simple: work should come to people in context.**

For citizens, the interface may be a portal, guided journey, messaging channel or voice experience. For staff, it may be a collaboration tool, workflow surface or operational command layer. What matters is that context, action and escalation arrive where the work happens.

If staff still have to copy, paste, chase, re-key or manually hand work between teams, the service has not been redesigned. It has just moved the coordination burden somewhere else.

Output of this Step

- A citizen channel strategy across web, voice, messaging and portal
- A staff work-surface strategy that reduces context switching
- A service journey showing where context, action and escalation appear

Common trap: Do not leave humans as the integration layer.

If staff are still copying, pasting, chasing updates and manually handing work between teams, the agent has not removed coordination burden. It has simply moved it.

01 02 03 04 05 06

05

CAPABILITY FIVE

Human handoff and observability

An agent should not rely on generic model recall when a service depends on current policy, procedural nuance or local interpretation. It must be grounded in trusted organisational knowledge: guidance, decision criteria, service standards, process rules and, where appropriate, case history.

This is what turns a fluent model into a useful worker. If the agent cannot reliably access the right knowledge, it may sound impressive while remaining operationally weak. It may generate plausible responses, but may not generate trust. In practice, this often matters more than model choice.

Output of this Step

- A human handoff model with defined escalation points
- Observability requirements for agent behaviour and outcomes
- A review loop for failure modes, complaints and edge cases
- Clear accountability for improving the handoff over time

What Good Looks Like

A well-designed agentic service does not begin with a general-purpose chatbot. It begins with a tightly bounded operational problem.

That is what turns a model into a real service.

A service is only ready to scale when these capabilities exist together. Remove one, and the experience becomes weaker, riskier or harder to trust.

The stack only works as a system. Trusted knowledge without action only advises. Action without handoff creates risk. Fragmented interfaces push coordination back onto people.

01 It is grounded in trusted knowledge.

02 It is equipped with meaningful action.

03 It is designed through composable capabilities.

04 It is presented through a consistent interface for work.

05 And it is governed by explicit human handoff and observability.



06 | Governance that Enables Scale

Designing control that Accelerates Delivery



“Good governance is not heavy. It is clear.”

Governance That Enables Scale

Designing Control That Accelerates Delivery

Governance should make agentic delivery easier, not slower.

When governance is weak, organisations hesitate. They stay trapped in pilots, blur accountability and struggle to move from experimentation to live service.

When governance is clear, teams move faster because the boundaries are understood, owners are visible and trade-offs can be made deliberately.

This matters with agents because they do not just generate content. They retrieve information, trigger actions, update records and shape operational outcomes. **That means governance must move beyond abstract AI principles and into service-level reality.**

The Real Question

The question is not whether your organisation has an AI policy.

It is whether it knows:

- Who owns the outcome?
- Where automation stops?
- How risk is handled?
- How live performance will be reviewed once the service is running?



Governance That Enables Scale

The Core Principles

Good governance accelerates delivery. Bad governance blocks it.

01 | Ownership

If technology owns the platform, policy owns the rules and operations own delivery, but no one owns the end-to-end outcome, the programme will stall.

A live agentic service needs:

- A named owner accountable for the service outcome
- A clear owner for policy or rules
- An operational owner for performance and iteration

These roles may sit in different parts of the organisation, but they must be aligned.

02 | Rules vs Judgement

Not every part of a service should be automated. Some work is rules-based. Some depends on human judgement. Most services contain both.

The discipline is to separate them explicitly.

- Let the agent handle high-volume, repeatable work
- Reserve human judgement for exceptions and sensitive decisions

This creates scale without pretending every decision can, or should, be automated.

03 | Proportional Risk

A status update, a password reset and a safeguarding decision should not carry the same governance burden.

Apply governance in proportion to service criticality.

- Low-risk, reversible tasks move quickly
- High-risk, consequential decisions require stronger controls

This allows progress without compromising safety.

04 | Auditability

If you cannot explain what happened, you are not ready to scale. A public service using agents should be able to show the audit trail.. A public service using agents should be able to show:

- What the user asked
- How the request was interpreted
- What knowledge or rule was used
- What action was taken
- Where human escalation occurred

This is what turns an interesting capability into an operationally credible one.

Multidisciplinary Delivery Practice

Shared outcomes. Clear ownership. Transparent metrics. Rapid escalation.

Governance That Enables Scale

What Good Looks Like

It makes ownership visible, separates rules from judgement, applies controls in proportion to risk and ensures every consequential action can be understood after the fact.

That is what allows an agent to move from pilot to trusted service.
Governance should be small enough to move, but senior enough to decide.

Aim for 5–10 named stakeholders across service ownership, policy, operations, digital, data, assurance and risk. The goal is not a bigger steering group; it is a small decision forum that can resolve trade-offs, unblock delivery and keep the service aligned to the outcome..

Named Ownership Model

Service owner, policy owner and operational owner clearly defined

Automation Boundary

Clear separation between rules-based work and human judgement

Risk Tier and Control Model

Governance proportional to service criticality

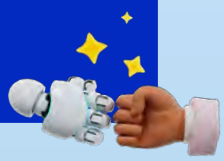
Auditability Requirements

Consequential actions visible, traceable and reviewable

Multidisciplinary Delivery Practice

Transformational services rarely sit neatly inside one team. They cut across policy, operations, digital, data and assurance. That means delivery depends on alignment without authority.

The answer is not a larger steering group. It is a small multidisciplinary delivery practice built around a shared service outcome, with clear ownership, transparent metrics and rapid escalation paths.
This is how governance becomes coordinated delivery, not delay.





Governance is not the barrier to scale. It is the enabler of trust.

07 | Measuring What Matters

Proving Service Transformation, Not Just AI Activity



“Build the baseline before you build the story.”

Measuring What Matters

Proving Service Transformation, Not Just AI Activity



One of the easiest ways to kill an agentic programme is to measure the wrong thing.

Many organisations still evaluate AI through the lens of novelty, adoption or user enthusiasm. They count prompts, track usage, monitor token consumption, celebrate time saved and assume value will follow. But token usage is not public value. Activity is not transformation.

Sometimes it does. But transformation demands more. If the ambition is transformation, those measures are not enough. Transformation is not proven by whether people like the tool. It is proven by whether the service performs differently.

If the metrics do not tell a service story, they are probably still telling a feature story.

Common trap: Do not mistake usage for value. High token consumption, prompt volume or adoption may show that people are experimenting. It does not prove that demand has fallen, access has improved or the service model has changed.





Start with the Service, Not the Tool

01 Fewer status calls

02 Faster triage

03 Lower cost per case

04 Fewer incomplete submissions

05 Reduced backlog

06 Improved first-contact resolution

07 Better access across time, language or channel

These are the measures that matter because they describe a change in the operating model, not just the user experience of the technology.

Measuring What Matters

Measure Across Four Layers

In practice, it helps to think about measurement in four layers



Adoption

Asks whether the capability is being used. Are people returning to it? Is demand being contained rather than escalated?



Service performance

Asks whether the service is becoming faster, clearer, more accessible or more reliable. Are repeat contacts falling? Are cases moving more quickly?



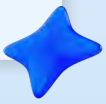
Economics

Asks whether cost per interaction is falling, expensive demand is shifting out of phone channels and staff effort is being released from routine work.



Trust and Quality

Asks whether the service is accurate, defensible, fair and suitable for scale. Are escalations appropriate? Are complaints falling? Is the service auditable?



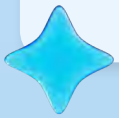
All four matter. But only the middle two tell you whether transformation is actually happening.

Build the Baseline Before You Build the Story

A surprising number of AI programmes struggle to prove value because they failed to established a clear baseline. If you do not know your current cost per case, repeat-contact rate, escalation volume or cycle time, then every success claim becomes anecdotal and every budget request becomes bargaining.

This is why the baseline work in the diagnostic matters. It gives you something to compare against. It allows you to distinguish between genuine improvement and background noise.

If the baseline is weak, the story will be weak.



Avoid Vanity Metrics

User satisfaction, prompt volume, number of automations deployed and self-reported time savings can all help explain local improvement. None of them, on their own, proves that the service model has changed.

A good measurement model makes that impossible. It forces the organisation to connect adoption to service performance, service performance to economics, and economics to strategic value. If the service is no faster, cheaper, clearer or more accessible, the programme has not yet delivered transformation – regardless of how impressive the demo looked.

Measuring What Matters

What Good Looks Like

Proving service transformation, not just AI activity

A strong measurement model connects activity to outcomes, outcomes to economics, and economics to public value.

Trust and Quality	Is the service accurate, defensible, fair and auditable? Are escalations appropriate? Is confidence improving?
Economics	Is cost per interaction falling? Is demand shifting out of expensive channels? Is staff effort being released from routine work?
Service performance	Is the service faster, clearer, more accessible and more reliable? Are repeat contacts falling? Are cases moving more quickly?
Adoption	Is the capacity being used? Are people returning to it? Is demand being contained rather than escalated?

Only the middle two prove transformation. **The other two determine whether it can scale.**

What this requires

Service Baseline	Four-Layer Model	Transformation Metrics	Review Cadence	Vanity Metric Filter
Cost, time, repeat contact, escalation and access clearly defined.	Adoption, service performance, economics and trust all measured.	A small set of agreed operational outcomes, linked to the service goal.	Regular measurement rhythm with named owners and escalation routes.	A clear rule that excludes soft or self-reported gains as proof of transformation.

The question is not whether the tool was used.
It is whether the service is now performing differently.

08 | Scaling Without Losing Control

Scale Patterns, Not One-Offs



“The goal is not to recreate every success from scratch. It is to identify the design patterns that can be reused across services.”

Scaling Without Losing Control

Scale Patterns, Not One-Offs

The next challenge is scale. This is where many organisations repeat the same mistake they made at the pilot stage. They scale enthusiasm instead of architecture. They multiply experiments instead of patterns. They add agents faster than they create coherence. The result is predictable: duplication, policy inconsistency, unmanaged cost and a new layer of fragmentation.

The discipline of scale is therefore not to increase the number of agents. It is to increase capability without increasing chaos. Scale is not the multiplication of agents. It is the multiplication of repeatable patterns.

Scale Patterns, Not One-Offs

A successful use case should not be treated as an isolated win. It should be treated as a repeatable pattern. **What made it work?**

- Workflow Shape
- Data Access Pattern
- Knowledge Structure
- Escalation Design
- Service Interface
- Ownership Model

These are the things that should be scaled. The goal is not to recreate every success from scratch. It is to identify the design patterns that can be reused across services. That is how an organisation moves from scattered use cases to a coherent operating model.



Scaling Without Losing Control

Scale Patterns, Not One-Offs

Scaling successfully requires two disciplines.

You must avoid bad scale patterns – and build the internal capability required to sustain good ones.

DISCIPLINE ONE

Avoid the Monolithic Agent

Do not scale by building one agent to do everything. Scale by orchestrating bounded capabilities.

Just like humans, Agents perform best when they are given controlled, relevant context, and a clear path to what 'good' looks like.

There is a natural temptation, once confidence grows, to design a single enterprise agent that can do everything.

This creates confusion rather than scale.

Monolithic agents have a 'context engineering' problem; that is, they have many rules to follow and many actions they could take; if they receive ambiguous inputs, there's a higher risk they execute the wrong task.

Instead, design discrete 'sub-agents'; if you need to scan a document for identity information, that's one sub agent. If you need to read the same document for eligibility assessment, that's a separate sub agent.

Agents like orchestrating above well defined sub-agents; quality and accuracy reduces the more context you add to a single agent.

Scale comes from orchestration, not from creating one or two agents to do everything.



DISCIPLINE TWO

Build Internal Capability

Do not scale by deepening dependency on external build teams. Scale by growing service-owned capability close to the work.

The beauty of building Agents compared to traditional IT software is instructions are added in natural language. We no longer need to turn flow charts or operating procedures in to code; we give it to the Agent and to do this for us.

The most critical success factors to your Agent implementation are your business people; your ops staff who know how to execute a service process correctly, or your policy teams across the intent of departmental policy.

Best practice is to co-build Agents with business and IT teams, with the business seeing Agents in to live service and iterating them to make them even better. The business defines and owns what good looks like; IT enables the plumbing to make it happen, like connecting to the right data sources.

Too much organisational change still relies on IT teams acting as the gatekeepers between business needs and live delivery. Agentic systems remove that barrier, with IT operating as governance not gatekeepers.

If every prompt tweak, workflow adjustment, or policy update still requires you to log a ticket with a separate build team, then your organisational capability hasn't grown. The target state is the business monitoring and optimising Agents, following the guardrails IT define.

It is important you align all your teams around this goal; an enduring, business driven capability governed through isecure guardrails. New technology requires new ways of working to fully realise benefits.

Scaling Without Losing Control

What Good Looks Like

A scaled agentic organisation does not look like the one with the most agents. It is the one with the clearest patterns, the strongest internal capability, the most coherent governance and the clearest service outcomes.

A mature scaling model includes:

Reusable patterns

A catalogue of patterns and agent components you can reuse across services.

Composable architecture

A composable design, not a monolithic one. Bounded capabilities, orchestrated deliberately.

Internal capability

A capability plan spanning policy, operations and digital – with named roles.

Coherent scale

A route to reduce dependency on external build resources over time.

Scaling roadmap

Built around patterns, not isolated pilots. Measured by service outcomes.

Scale without coherence is just a faster route to fragmentation.



09 |

Conclusion



Conclusion

What Public Services Must Protect and Amplify

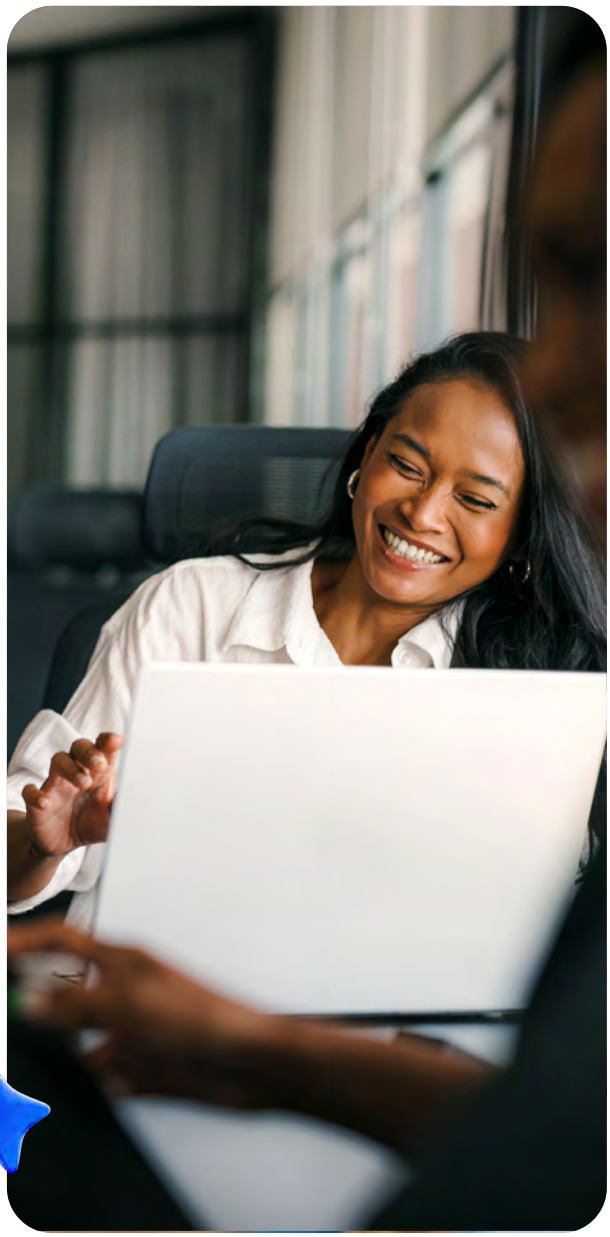
Public services have only two enduring assets: their people and their institutional knowledge.

That is the central argument of this handbook.

To truly unlock the value of agentic AI in public services, AI must strengthen both.

It should make institutional knowledge more accessible – to the public, so that citizens do not need to call, wait, chase and repeat; and to staff, so that they can act with greater speed, confidence and consistency.

At the same time, it should remove low-value work from people, so that human effort is focused where it creates the most value: complex cases, difficult decisions, vulnerable citizens and moments requiring trust.



That is the standard.

If an AI deployment does not make knowledge more useful, or does not free people to do more meaningful work, it is unlikely to be transformational.

The organisations that succeed will not be the ones with the most pilots or the loudest announcements. They will be the ones that unlock knowledge at scale, redesign access around real human need and redeploy people to higher-value work.

- The model will change.
- The market will change.
- The technology will continue to evolve.

What must endure is simpler than that: trusted knowledge, capable people and services designed to make the best use of both.

That is the task.

And that is where transformation begins.

Conclusion

The Transformation Blueprint

01	Start with the service outcome	Define the service in citizen or frontline terms. Do not start with the technology. Start with the result the service exists to create and the human experience it should improve.
02	Find the real opportunity	Surface friction, failure demand and cost concentration. Look for repeat contact, delays, broken handoffs, access barriers and high-volume pain points. That is where transformation becomes possible.
03	Design for transformation, not augmentation	Do not ask how AI can make existing work faster. Ask where demand can be removed, where access can improve and where the service itself can be redesigned.
04	Build the capability stack	A real service is not built from a model alone. It requires trusted knowledge, meaningful action, composable design, a consistent interface, and explicit human handoff with observability.
05	Govern and measure deliberately	Make ownership clear. Separate rules from judgement. Apply controls in proportion to risk. Measure service performance, economics and trust not just usage or sentiment.
06	Scale patterns, not pilots	Do not multiply experiments. Identify what worked, codify it, reuse it and build internal capability around it. That is how isolated success becomes a repeatable operating model.

Start with the service.
Build the capability.
Prove the value.
Scale the pattern.



The information provided in this report is strictly for the convenience of our customers and is for general informational purposes only. Publication by Salesforce, Inc. does not constitute an endorsement. Salesforce.com does not warrant the accuracy or completeness of any information, text, graphics, links, or other items contained within this guide. Salesforce.com does not guarantee you will achieve any specific results if you follow any advice in the report. It may be advisable for you to consult with a professional such as a lawyer, accountant, architect, business advisor, or professional engineer to get specific advice that applies to your specific situation.

© Copyright 2026, Salesforce, Inc. All rights reserved.

Sources

01 **MIT NANDA, The GenAI Divide: State of AI in Business 2025, Jul 2025.**
~5% of custom genAI pilots reach production. MIT-linked, not peer-reviewed.

02 **DSIT / DCMS, AI Skills for Life and Work:**
General Public Survey Findings, 28 Jan 2026. 73% of UK public had used AI in the past month.

03 **UKRI, AI Research and Innovation Strategic Framework, 19 Feb 2026.**
UKRI committed >£1.6 bn directly to AI research and innovation.

04 **DSIT, AI Opportunities Action Plan: One Year On, 29 Jan 2026.**
UKRI committed >£1.6 bn directly to AI research and innovation.

05 **Local Government Association, State of the Sector: AI – 2025 Update, 2 Jun 2025.**
95% of responding English councils using or exploring AI (snapshot, not census).

06 **Alan Turing Institute / DfT / ONS, Mapping the Potential: Generative AI and Public Sector Work, 2025.**
~41% of public-sector time theoretically supportable by generative AI.

07 **Ipsos / Tony Blair Institute, What the UK Thinks About AI, Sep 2025.**
Lack of trust a major barrier; 38% cited it. Verified via Ipsos summary.

08 **GDS, M365 Copilot Experiment: Cross-Government Findings Report, 2 Jun 2025.**
20,000 civil servants across 12 organisations; avg 26 min/day saved.

09 **Salesforce, Agentforce Customer Zero.**
Agentforce on Salesforce Help resolves >68% of conversations; >2 m interactions. Salesforce-reported.

10 **Salesforce UK Newsroom, UK Police Forces Transform Citizen Support with 'Bobbi', 4 Dec 2025.**
UK police forces deploying live, workflow-integrated digital agents.

11 **DHSC / NHS England, AI to be trialled at unprecedented scale across NHS screening, 22 Sep 2025.**
Supporting evidence for AI-assisted diagnostics scaling in NHS.

SOURCE NOTE

Figures drawn from publicly available reports, official publications and case-study material. Government and research sources used wherever possible. Salesforce examples represent Salesforce-reported customer or internal deployment metrics.