

AI IN GOVERNMENT

The Practitioner's Playbook

How to build a high-impact AI team inside the state?

Lessons from the UK's Incubator for AI (i.AI)

A first-principles guide for governments building sovereign AI capability

Authors: Alexander De Ville (lead author), with Alex Jones, Liam Wilkinson, Sharif Kazemi, and Shahrukh Wani.

Note: This is a live draft intended for discussion and feedback and is subject to revision. The views expressed are those of the authors and do not necessarily reflect those of the organisations with which they are affiliated. The authors are grateful to Jeremy Ng and Harshali Ranjan for their feedback.

[Foreword](#)

[Why This Moment Is Different](#)

[The Cost Evidence for In-House Capability](#)

[What in-house actually costs](#)

[What the contractor alternative costs](#)

[The CAPEX trap](#)

[Five outsourcing failures](#)

[The Strategic Triangle](#)

[Wedge Cases vs Edge Cases](#)

[Big Technical Questions - What Government Actually Needs](#)

[Chapter 3: How i.AI Actually Operated - and Where It Is Heading](#)

[What Was Actually Done: Productivity + Big Bets](#)

[The Full Product Portfolio - What Has Been Built](#)

[The Knowledge Hub - A Diffusion Mechanism](#)

[Six Stages with Kill Criteria](#)

[The Assurance Toolkit](#)

[OKRs - High Alignment, High Autonomy](#)

[Chapter 5: Talent - The Single Biggest Unlock](#)

[The Talent Flywheel](#)

[What Destroys the Flywheel](#)

[Chapter 6: Incubation in Practice - Four Case Studies](#)

[What Good Incubation Looks Like](#)

[Chapter 7: Scaling - From Pilot to Production](#)

[The Question the Field Keeps Avoiding](#)

[The Three Structural Barriers to scaling](#)

[Three Scaling Models - and What i.AI Actually Tried](#)

[The Honest Scaling Heuristic: No Single Rule](#)

[Infrastructure as the Scaling Ceiling](#)

[Chapter 8: Build, Buy, or Partner](#)

[The Framework](#)

[Using New Procurement Powers](#)

[Chapter 9: International Context - What Others Have Built](#)

[Singapore GovTech: The Most Comparable Operation at Scale](#)

[For Middle Powers and Developing Country Contexts: Ukraine and the Open-Source Imperative](#)

[The Middle Powers Argument: What the TBI Analysis Shows](#)

[Chapter 10: What went wrong](#)

[Chapter 11: The First 90 Days](#)

[Days 1–30: Foundation](#)

[Days 30–60: First Project](#)

[Days 60–90: Demonstrate and Learn](#)

[Year One Milestones](#)

[Appendix A: Decision Checklists](#)

[Project Selection Checklist](#)

[Stage Gate Pre-Requirements Summary](#)

Foreword

Governments everywhere are asking the same question: how do we actually do AI? Not in the sense of which LLM to buy, but in the deeper sense - how do we build the institutional capability to deploy AI in ways that genuinely change what government can do?

The UK's Incubator for AI (i.AI) began as a spin-off from the Number 10 Data Science team. Three or four people, with a mandate to figure out what AI could do for government, and the principle that tiny teams moving quickly can deliver big things. Two years later, the team has grown to around 60, bringing external AI talent into government alongside some of the best civil servants, and has built, shipped, and published research and tools that span consultation analysis, planning data, education, and healthcare.

This playbook is an attempt to codify what was learned from that work - honestly. Not just the successes, but the failures, the pivots, and the things that are still unresolved. i.AI's willingness to say publicly that a flagship health project should not proceed to deployment, and to publish exactly why, is rare in government. That spirit runs through every chapter.

A note on scope: this playbook focuses on substantive AI implementation - building systems that change how government works. It is not about buying productivity tools, writing AI strategies, or running awareness campaigns. The approach throughout has been to learn by doing: try things, build them, attempt to deliver a policy the government wants to achieve in the most ambitious way possible, see where you fall short, and iterate again.

Who this is for

This playbook is written for practitioners - technical leads, product managers, engineers inside government - and for the ministers and senior officials without whose sponsorship nothing happens. You need all both audiences reading it together, because neither can succeed without the other. It is also written for international governments, development partners and funders learning from the UK experience.

Executive Summary

The UK Prime Minister cited £45 billion per year in unrealised productivity savings from digital transformation in March 2025. The GDS Copilot trial across 20,000 civil servants saved the equivalent of 1,130 person-years. Yet 55% of the UK government's digital budget goes to external contractors. The gap between potential and delivery is not a technology problem, it is an organisational, talent, and delivery problem.

The UK's Incubator for AI (i.AI) was created to close that gap. It started with 3–4 people as a spinoff from the Number 10 data science team. Within two years it grew to around 60 people, attracting talent from frontier AI labs, building products that are now in use. The model: tiny teams, moving quickly, building real things, publishing honestly - including the failures.

i.AI's product portfolio demonstrates what this model produces: Consult (tested on 26 live consultations, saving thousands of admin hours and over half a million pounds, with the potential to save 75,000 civil service days annually); Extract (digitising planning data, available to all local authorities from Spring 2026); and MedGuard (the first LLM medication

safety evaluation on real NHS data, sensitivity 100%, specificity 83.1%, published as academic research).

£45bn Unrealised annual productivity savings (State of Digital Govt Review, 2025)	55% UK digital budget spent on external contractors, not internal capability	1.83× Average contractor premium over equivalent permanent staff (POGO, US)
---	--	---

Theme	Key Finding	Playbook Chapter
Operating Model	Two tracks in practice: productivity tools and big bets, held together by a disciplined product lifecycle.	Chapter 3
Product Lifecycle	Six stages: Scoping→Incubation→Alpha→Beta→Scaling→Graduated. Stage gates with kill criteria at every transition.	Chapter 4
Cost	In-house capability is substantially cheaper than the contractor alternative at equivalent quality - but requires sustained political will to hire permanently rather than procure.	Chapter 1
Talent	Talent flywheel: quality attracts quality. Hackathons as the primary validation mechanism.	Chapter 5
Scaling	The Knowledge Hub is the cross-government diffusion mechanism, alongside open-sourcing and procurement-embedded deployment.	Chapter 8
International	Ukraine's Diia: 23M users, world's first national AI agent (Diia.AI, 2025). Singapore AIBots: 40K users in 6 months. Estonia: 12-person team, €13M. TBI framework: build the adaptation layer, not a national LLM.	Chapter 9

Chapter 1: The Case for Action

Why This Moment Is Different

Previous digital transformations were about making existing government processes faster and cheaper. AI can do that, but it also makes it possible to reimagine what those processes are. The i.AI experience shows this at a practical level: Consult does not just process consultation responses faster - it changes what consultations can ask, how many responses can be gathered, and whether meaningful democratic engagement is possible at scale. Extract does not just speed up planning decisions - it creates a digital dataset of planning precedents that has never previously existed.

There is also a temporal dimension that is unprecedented. Previous digital transformations happened after the technology had matured in the private sector. AI is different: the frontier is moving in real time, and government needs to develop its own capability to keep pace rather than waiting ten years to implement what was proved elsewhere in 2023.

The Cost Evidence for In-House Capability

The most common objection to building in-house AI capability is cost. The evidence consistently points the other way: in-house capability is substantially cheaper than the contractor alternative - and unlike contractor arrangements, it compounds rather than resetting to zero at contract end.

What in-house actually costs

i.AI draws on the government's published DDAT (Digital, Data, and Technology) pay framework, with London weighting. A senior AI engineer or data scientist in permanent government employment costs a fraction of what an equivalent contractor through a prime supplier charges per day - and that permanent employee compounds domain knowledge year on year. A contractor resets the knowledge clock when the engagement ends.

What the contractor alternative costs

Indicative rates for a senior developer through a government prime supplier are £800–£1,200 per day. A Big Four consultant on an AI project: £900–£1,400 per day. At 230 working days per year, that quickly adds up. And this does not include the management overhead of procurement, the knowledge loss at contract end, or the premium on repeat engagements when the expertise walks out the door.

Published evidence supports the structural case. POGO (US, 2011) found contractors cost federal agencies on average 1.83× equivalent permanent staff - and in over a third of roles studied, more than double. The UK NAO's November 2025 report (HC 1381) documented £1.36 billion in central government consultancy spend in 2022–23, with the MCA reporting £5.3 billion across the wider public sector in 2023. The NAO found the government 'does not have consistent data on how much individual departments spend on consultants' - suggesting these figures understate the real cost.

£1.36bn UK central govt consultant spend 2022-23 (NAO, Nov 2025)	1.83× Average contractor premium over equivalent permanent staff (POGO)	55% UK digital budget going to external contractors, not internal capability
---	---	--

The CAPEX trap

Government budget structures systematically favour outsourcing. A one-off consultancy contract is typically categorised as CAPEX (capital expenditure); permanent salaries are mostly OPEX (operational expenditure). On annual budget cycles, the CAPEX option consistently appears cheaper - even when it is dramatically more expensive over three to five years and builds no lasting capability. The NAO has identified this as a structural problem. The same structural logic - outsourcing as project spend, salaries as a permanent liability - has produced equivalent distortions in Australia, Canada, and across the EU, where governments have quietly built 'shadow workforces' of contractors that dwarf the civil service headcount they were designed to circumvent. HM Treasury had a target to reduce consultancy spend by £550 million in 2024–25.

Five outsourcing failures

Project	Budget → Final Cost	Multiplier
NHS NPfIT (UK)	£6.2bn → £10.1bn+	1.6×
Phoenix Pay System (Canada)	CAD \$309m → \$5.1bn+	~16.5×
Healthcare.gov (US)	\$93.7m → \$300m–\$1.7bn	3–18×
Universal Credit IT (UK)	£2bn → £15.8bn whole-life	7.9×
Robodebt (Australia)	Projected \$1.7bn savings → \$2.4bn+ redress	Net catastrophe

In each case, government had insufficient internal capability to specify what it needed, evaluate whether it was getting it, or challenge the contractor when things went wrong.

Chapter 2: Selecting the Right Problems

The Strategic Triangle

Every i.AI project must satisfy three conditions simultaneously. Projects that fail one condition - even if they succeed on the other two - are poor investments.

Politically Salient Aligns with ministerial priorities. Gets attention, resources, and air cover when things get hard.	Technically Challenging Interesting enough to attract and retain top talent. Pushes technical boundaries. Creates learning.	Appropriate for AI AI meaningfully improves outcomes vs alternatives. The problem has the properties AI is suited to solve.
--	---	---

Wedge Cases vs Edge Cases

Wedge cases are chokepoints in high-volume processes where AI changes outcomes for large numbers of people. Nearly 70% of planning applications are simple householder applications. A large share of applications arrive invalid, delaying substantive review by weeks. These are wedge cases: high volume, structured criteria, AI-appropriate.

Edge cases are technically interesting problems at the margins. They fascinate data scientists. They are almost always the wrong thing to work on first. The rare heritage-sensitive planning application is an edge case. A complex multi-drug interaction in an unusual patient is an edge case. Do not start there.

Big Technical Questions - What Government Actually Needs

The technical capability questions below are the ones the government genuinely needs frontier AI labs to answer. These are not the kind of questions that appear in strategy documents - they are the engineering questions that determine whether the products work in production.

The government has a target to build 1.5 million new homes. Nearly 70% of planning applications are simple householder applications, yet many residents wait months for what should be straightforward decisions. The ambition is to map the underlying data of the planning system, currently stored in paper documents and PDFs across 300+ local authorities, to enable automation of the simplest, highest-volume decisions.

TECHNICAL QUESTIONS FOR FRONTIER LABS::

- Object detection and image segmentation: extracting precise measurements, scales, and specifications from architectural drawings submitted at different scales and orientations
- Rank and reason over legislation and policy to arrive at legally sound conclusions with auditable reasoning chains - essential for automated decisions that must be legally defensible
- Consistent spatial reasoning across different coordinate systems and drawing conventions
- Complex geometric relationships and temporal changes (e.g. a building extension that overshadows a neighbour's garden at specific times of year - a genuine grounds for refusal)
- Verification mechanisms for agentic workflows: an auditable chain of thought that aligns with what an agent actually does (not just what it claims to do)

CURRENT STATUS: Extract converts planning documents to structured digital data (launched at London Tech Week 2025; available to all local authorities from Spring 2026). The MHCLG Augmented Planning Decisions programme, procured through a Competitive Flexible Procedure under the Procurement Act 2023 and contracted Google Cloud with Google DeepMind and delivery partner Faculty, to trial AI-assisted assessment with Barnet, Camden and Dorset councils, aiming to halve decision times on householder applications.

WHY THIS IS HARD: Planning data is stored in paper across ~300 local authorities with incompatible systems, using drawings submitted at multiple scales and orientations, referencing legislation across multiple national and local policy frameworks. Much of England's historical planning data has never been digitised, which means most local authorities cannot modernise their planning processes until the data layer is built.

The NHS delivers 370 million GP appointments annually. Delayed discharges affect approximately 14,000 patients daily, at a cost of over £400 per bed per day. Two of the most promising AI opportunities are (1) Rapid patient discharge - agentic solutions coordinating across social care beds, prescriptions, and transportation; (2) Doctor in your pocket - smart triage for secondary care, including appointment scheduling and avoiding unnecessary primary care attendance.

TECHNICAL QUESTIONS FOR FRONTIER LABS::

- Strategies for deploying and scaling AI solutions across the NHS's fragmented IT landscape
- Safeguards and governance mechanisms for AI agents acting in high-risk clinical domains - including how to handle uncertainty, when to escalate, and what constitutes an unacceptable false positive rate
- The specific capabilities needed for rapid discharge coordination: reading discharge summaries, identifying available social care placements, and coordinating transport across multiple systems with different APIs

MEDGUARD: i.AI's MedGuard project (arXiv 2512.21127, published December 24, 2025) is the first evaluation of an LLM-based medication safety review system on real NHS primary care data. It uses a population-scale EHR of 2,125,549 adults in NHS Cheshire and Merseyside. After data quality exclusions: 277 patients evaluated by expert clinician.

Results: Sensitivity 100% [95% CI 98.2–100], Specificity 83.1% [95% CI 72.7–90.1]. The system correctly identifies whether a clinical issue is present with perfect recall. But 17% of false positives means one in six patients flagged as needing intervention does not. The failure mode analysis (published in the paper) identified five structural ways LLMs fail in clinical settings.

This paper is significant beyond its findings. Only 5–8% of clinical AI studies use real patient data (vs synthetic or curated benchmark data). The published methodology - a three-level hierarchical evaluation framework - is the contribution the NHS 10-year plan requires for moving 'AI seamlessly integrated into clinical pathways.' Paper code: <https://github.com/i-dot-ai/medguard-paper>

The UK employs 470,000 teachers. One-to-one tutoring accelerates a child's learning by approximately five months, but access is deeply unequal. The government's education AI programme focuses on: personal AI tutors, AI marking and feedback, and benchmarking datasets from real-world deployment (described as 'HealthBench for Education').

TECHNICAL QUESTIONS FOR FRONTIER LABS::

- How are labs adapting AI models to align with specific pedagogical approaches - not engagement-maximising commercial designs but evidence-based pedagogy?
- What are the internal benchmarks for AI in long-term tutoring systems? How do they compare to previous Intelligent Tutoring Systems?
- Safety benchmarks for long-term relationships with children - in the classroom and elsewhere
- VLM performance in noisy educational environments: understanding handwritten mathematics and science, including spatial and visual reasoning
- Avoiding sycophancy and correcting users when there may be more than one valid answer - a critical pedagogical challenge
- Image generation for accurate scientific and mathematical diagrams

CURRENT STATUS: The DfE announcement (January 2026) committed to AI tutoring tools that could support up to 450,000 disadvantaged secondary pupils by the end of 2027, with trials in schools beginning later in 2026. Eight companies have been selected to co-create tools with teachers, and robust benchmarks are being developed so that parents and teachers can be confident the tools are high quality, reliable and safe.

WHY THIS IS DIFFERENT FROM COMMERCIAL AI EDUCATION: Commercial AI tutoring tools are often optimised for engagement rather than measured learning outcomes. The government must take a fundamentally different approach - building common infrastructure for what a pupil should be learning, what they have already learnt, and how they are being taught. This context layer is the technical innovation, not the underlying model.

Chapter 3: How i.AI Actually Operated - and Where It Is Heading

What Was Actually Done: Productivity + Big Bets

The approach that generated the case studies in this playbook - Consult, Extract, MedGuard, the AI Education programme - was two parallel tracks of work, held together by heavy experimentation and a disciplined product lifecycle.

Productivity tools	Improvements to how civil servants work day-to-day: Redbox (document analysis and chat for Cabinet Office, No10 and DSIT, built before any enterprise LLM solution was available in government), Minute (AI transcription, piloted across 22 local authorities and now being adapted by MHCLG as Local Transcribe), Lex (intelligent search across UK legislation back to 1267, available as an MCP and open-sourced), Parlex (parliamentary intelligence across Hansard, Parliamentary Questions and committee materials), Justice Transcribe (the Ministry of Justice's fork of Minute for probation meetings), and consultation analysis (Consult)..
Big bets	Higher-risk, longer-horizon projects where AI could fundamentally transform public services: Extract (digitising planning data), the AI Education programme (AI tutors for 450,000 pupils), MedGuard (NHS prescription safety - pivoted to published research when deployment proved premature). These had uncertain technical paths, required deep departmental partnerships, and generated both the most significant impact claims and the most significant failures.

What held the two tracks together was not a formal governance structure but the product lifecycle discipline: clear stage gates, explicit kill criteria, regular demos to users, and a culture of failing fast and publishing honestly. There was no single rule for scaling - it depended on the product, and as the team gained experience, they got more effective at scaling it to more and more people more cheaply.'

The characterisation

i.AI was fundamentally an experiment in whether a small, high-quality in-house team could build and deploy AI faster and better than the alternatives. The answer across two years: yes, on the building side, with significant lessons still being learned on the scaling and diffusion side. The operating model emerged from experience and was codified retrospectively - not designed in advance.

The Full Product Portfolio - What Has Been Built

Product	Track	Stage / Status	Value Case
Redbox	Productivity	Open-sourced.	Document analysis and chat for Cabinet Office and No10. Built before enterprise LLM solutions available in UK government.
Minute (AI transcription)	Productivity	Piloted with 22 local authorities and Ministerial Private Offices; open-sourced; being adapted by MHCLG as Local Transcribe	Meeting and interview transcription. Forked by the Ministry of Justice as Justice Transcribe for probation meetings.
Lex (legislation search)	Productivity	Live; API publicly available; open-sourced	Semantic search across UK legislation back to 1267, drawing on 1.5 million sections of legislation and 63,000 court cases, accessible via MCP
Consult	Productivity	Testing with live consultations; wider rollout planned for 2027	Potential to save 75,000 civil service days annually (£20M equivalent). Tested on 26 live consultations; published evaluations with the Scottish Government, Defra and DWP; F1 scores of 0.76 to 0.82.
Parlex	Productivity	Piloted with five government teams (November 2025)	Parliamentary research assistant across Hansard transcripts, Parliamentary Questions and Select Committee materials, with the potential to save thousands of hours across the civil service.

Extract	Big Bet	Live; available to local authorities from Spring 2026	Digitising historical planning data: trials digitised planning records including maps in about three minutes versus 1–2 hours manually. Feeds the Augmented Planning Decisions programme, which aims to halve decision times for householder applications.
AI Education	Big Bet	Active; co-creation with eight Pioneer companies; trials from 2026	Robust benchmarks in development so parents and teachers can be confident tools are high quality and safe. Target: up to 450,000 disadvantaged pupils by end 2027.
MedGuard	Big Bet	Pivoted to published research	Sensitivity 100%, Specificity 83.1% on 2.1M-patient NHS EHR. arXiv:2512.21127. First such evaluation on real NHS data.

The Knowledge Hub - A Diffusion Mechanism

One horizontal infrastructure was built, termed the AI Knowledge Hub (ai.gov.uk/knowledge-hub). It has grown steadily, with new features including a Prompt Library being added over time. The ambition was to move from 'static guidance' to a true toolkit: a place where government teams can log in, find tools, lift and shift them, and actually use them. It contains seven capability categories: Find Information; Analyse Data and Text; Collect and Clean Data; Create and Review Documents; Triage and Route Information; Run Meetings and Appointments; Digitise Documents.

The Knowledge Hub also reflects how i.AI actually scales: not by trying to deploy its own tools everywhere, but by coaching other departments, de-risking their innovation ('if you have a fantastic idea about an AI system, we will help you build the prototype and share it') and building the shared infrastructure that prevents 300+ organisations from independently building the same tools

Chapter 4: The Product Lifecycle - From Idea to Production

Six Stages with Kill Criteria

Every i.AI product goes through six stages. The critical discipline is the stage gate: a decision session that determines whether a product progresses or is killed. Not every product should progress. A kill at Incubation stage is not a failure - it is the model working correctly.

Stage	Purpose	Pre-Gate Requirements	Exit: Graduate or Kill
 Scoping	Define the problem and come up with AI solution hypotheses	Clear problem statement. Stakeholders identified. Build vs buy assessed.	Graduate to Incubation if: real user need, clear value case, data accessible, problem AI-suited. Kill if: no clear user, private sector already solving it well, no data path.
 Incubation	Run technical experiments. Address uncertainties. Prove the model works.	Completed project proposal. Streamlit demo or prototype. Model tested. Theory of Change drafted.	Graduate to Alpha if: AI technically feasible, user need confirmed, stakeholder buy-in. Kill if: model not good enough, user need not confirmed, viability blocked.
 Alpha	Build and technically evaluate solution. Focus on lean/fast towards process evaluation.	Architecture designs. User journey proven. Reference customers identified. Model accuracy sufficient.	Graduate to Beta if: product works, engaged reference users, evidence of product-market fit. Kill if: accuracy insufficient, no reference customers, insurmountable technical blockers.
 Beta	Productionise the solution and test with real users. Impact evaluation plan.	Reference customers achieved. Quality metrics met. Impact evaluation plan in place.	Graduate to Scaling if: real users finding it useful, impact demonstrated, infrastructure proven. Kill if: engagement does not materialise,

			savings case does not hold.
 Scaling	Launch and rollout. Begin handover to permanent team.	Positive RCT or equivalent. Rollout plan. Business case for funding.	Graduate to Graduated when: product self-sustaining, permanent team in place, i.AI no longer needed.
 Graduated	Formal handover. Product self-sustaining without i.AI support.	Long-term funding strategy. Handover completed.	Ongoing governance check-in. Case study published.

The Assurance Toolkit

i.AI has built an assurance toolkit that makes it easier for product teams to test their products, fail fast, and move forward with confidence. The key tools at each stage:

Scoping	Project Scoping Proposal. Customer letter (imagine the world where the product has been built successfully - what does the grateful user write to thank you?). Commercial offering scan to check the private sector is not already solving it.
Incubation	Theory of Change - why and how the product will deliver impact. Reference customers - real users using the tool in earnest, willing to advocate for it. RAID log. Pre-mortem / advance hindsight: assume the product has failed, then ask why.
Alpha	Internal code review. Unit tests (pytest). Custom evaluation benchmarks - ground truth datasets with expected inputs and outputs, automated LLM evaluation using tools like Inspect.
Beta	Red teaming - team members or external teams taking on the role of malicious users. Penetration testing. DPIA (Data Protection Impact Assessment). Equality Impact Assessment. Algorithmic Transparency Reporting Standard.
Scaling	Load testing - what happens when user numbers are significantly higher? Summative evaluation (RCT or A/B test). Lessons learned session.

OKRs - High Alignment, High Autonomy

i.AI uses OKRs (Objectives and Key Results) to maintain clarity on goals while empowering teams to figure out how to achieve them. The target state is high alignment + high autonomy: teams focus on what problem to solve, teams figure out how to solve it.

The OKR cycle is quarterly and tied to the quarterly incubation cohort reviews. Team-level OKRs set the overall direction. Product-level OKRs define what a specific product is trying to achieve and how success will be measured. The US Healthcare.gov team's recovery provides a useful reference: a single objective ('Fix the website for the vast majority of people'), with four measurable key results (70% get through; 1-second response; 1% error rate; 99% uptime).

Key product team principles

1. Fail fast. Realise what isn't going to work sooner. Move on before investing too much. 2. Solve real user problems. Get the whole team close to users early. 3. Clear impact. Be good at killing projects that don't have a clear path to impact. 4. Empower teams. The best ideas come from the team working on the problem. 5. Low burden and admin. Speed up the innovation cycle - give teams clarity and support, not process.

Chapter 5: Talent - The Single Biggest Unlock

The Talent Flywheel

The talent flywheel: quality people attract other quality people. i.AI recruits reported being astounded by the quality of the people in the team. But the flywheel spins both ways. Teams dependent on consultancies cannot recruit; unable to recruit, they become more dependent on consultancies.

The talent model requires permanent engineers who compound domain knowledge year on year

What Destroys the Flywheel

- **Bureaucratic obstruction.** IT access taking months, mandatory training irrelevant to the work, procurement processes six times longer than necessary. These are individually minor and collectively devastating. Fighting this on behalf of the team is the leader's job.
- **Working on the wrong problems.** Excellent people need technically challenging work. A team stuck on a pilot for 18 months without shipping anything real will lose its best people.
- **No career path for technical specialists.** Government career structures were not designed for people who want to advance through technical excellence rather than management. Create genuine technical career ladders or lose your best engineers to industry.
- **Visible consultant dependency.** One clear way to signal that the team does not build and deliver tools in-house is to be obviously dependent on external contractors for delivery. This undermines recruitment.

Chapter 6: Incubation in Practice - Four Case Studies

What Good Incubation Looks Like

The best way to understand what good incubation looks like is through specific examples. Different i.AI products illustrate different aspects of the incubation discipline: speed to working prototype, evaluation-first design, honest assessment of failure, and data-as-prerequisite thinking.

i.AI CASE STUDY | Consult - Evaluation-First, Human-in-the-Loop by Design

PROBLEM: Government consultations take months to analyse, consume 75,000 days of civil service time annually, and cost an estimated £20 million per year - often outsourced to external consultancies at significant cost. A team of two or three civil servants might spend months going through 20,000 survey responses. This is precisely the kind of high-volume, repetitive pattern-matching task where AI is well-suited.

WHAT WAS BUILT: A human-in-the-loop system with intentional friction. The friction is not a bug - it is the design principle. The system extracts themes using LLMs; the analyst reviews, refines, and signs off. The dashboard then shows theme prevalence across demographics and response types.

EVALUATION (agreed before development, not after): Consult has been tested on 26 live consultations, with published evaluations alongside the Scottish Government, Defra and DWP, including the Independent Water Commission Call for Evidence and the Scottish Government consultation on non-surgical cosmetic procedures. F1 scores ranged from 0.76 to 0.82. And a more important finding: on the Water Commission evaluation, Consult's agreement with each reviewer group (0.79 and 0.82) was higher than the reviewer groups' agreement with each other (0.74).

SOCIAL READINESS: i.AI partnered with NESTA to earn the first-ever AI Social Readiness Label for a UK government product. This tested acceptability with the general public, which was crucial. The findings shaped design decisions about transparency and oversight.

Initial testing suggested Consult is around 400 times cheaper than existing analysis methods, processing thousands of responses in hours. Tests on 26 live consultations saved thousands of admin hours and over half a million pounds. At scale: 75,000 civil service days annually (£20M equivalent).

THE BIGGER AMBITION: Consult is not just about efficiency. 'Analysing all of these consultations is making a lot of people in government realise that maybe we need to change the way we engage with the general public. In a few years' time, consultations themselves might be completely reimagined.' The deeper ambition is participatory democracy: not just a faster processing tool but a foundation for more meaningful public engagement.

i.AI CASE STUDY | Extract - Data Infrastructure as the Product

PROBLEM: Most English local authorities cannot modernise their planning processes because their historical planning data has never been digitised. Only London and a few major cities have digitised planning data. The overwhelming majority of England does not.

WHAT IT DOES: Extract converts planning documents - including decades of paper records, scanned PDFs, and Article 4 directions - into structured digital data on an interactive map. OCR + layout analysis, geospatial entity extraction, document type classification. Jobs run asynchronously; status dashboard shows completed, awaiting, and failed.

THE INSIGHT: This is not an AI planning tool - it is data infrastructure that makes AI planning tools possible. The AI application (Augmented Planning Decisions) depends entirely on this data layer.

SCALING APPROACH: Rather than central operation or open-sourcing, Extract is being scaled through MHCLG: made available to all local authorities from Spring 2026, with the related Augmented Planning Decisions programme procured under the Procurement Act 2023 and contracted to Google Cloud, Google DeepMind and delivery partner Faculty. This embeds ownership at the receiving end from the start.

OUTCOME: Launched at London Tech Week by PM Keir Starmer, with a commitment to make Extract available to all local authorities by Spring 2026. The related Augmented Planning Decisions programme aims to halve decision times for householder applications, which make up nearly 70% of all applications.

i.AI CASE STUDY | MedGuard - Three Pivots, Honest Failure, and Published Research

Full citation: Normand O, Borsi E, Fruin M, Walker LE, et al. 'A Real-World Evaluation of LLM Medication Safety Reviews in NHS Primary Care.' arXiv:2512.21127v1 [cs.AI], 24 December 2025. Code: <https://github.com/i-dot-ai/medguard-paper>

PROBLEM: 22,000 avoidable deaths annually linked to prescription issues - polypharmacy risks compound with every additional medication an ageing patient takes. 1.4 million excess bed days. £1 billion NHS cost.

PIVOT 1 - SYNTHETIC DATA PROTOTYPE: The team 'quite naively thought we could take a large language model, pop in some patient information, and out the other side we would get really useful outputs for pharmacists.' The prototype, enhanced with BNF and NICE Guidelines tool-calling, passed a pharmacy exam using synthetic data and was demonstrated to early pharmacy and NHS partners. Their verdict was clear: the outputs were far too high level, and the system assumed the prescription was causing the problem when the true cause might be visible only in a full GP record. Pivot.

PIVOT 2 - PHARMACY PARTNERSHIP DATA: Testing on real pharmacy data revealed the same structural problem. The model was making assumption after assumption. Pivot.

PIVOT 3 - 2.1M-PATIENT EHR (NHS Cheshire and Merseyside): The ICB had over 2 million patient records. i.AI set up a secure research data environment - something only 5–8% of clinical AI studies achieve - and tested the same system on real GP data at population scale. 277 patients evaluated by expert clinician.

RESULTS: Sensitivity 100% [95% CI 98.2–100]; Specificity 83.1% [95% CI 72.7–90.1]. The model caught every genuine issue. But 17% of flagged patients needed no intervention, and the system correctly identified all issues and interventions in only 46.9% of patients [95% CI 41.1–52.8].

FIVE FAILURE MODES (published in the paper):

1. Acts decisively when uncertainty should prompt more information-gathering (n=51)
2. Applies standardised guidelines without adjusting for individual patient context (n=49)
3. Understands formal documentation but misunderstands how healthcare is actually delivered (n=30)
4. Produces sensible clinical reasoning but with unreliable factual content (n=25)
5. Identifies correct clinical endpoints but recommends unsafe pathways to reach them (n=23)

THE PIVOT TO RESEARCH: Rather than deploying a flawed system or quietly abandoning the project, i.AI characterised all failure modes and published them. The findings are now informing NHS data infrastructure redesign.

i.AI CASE STUDY | AI Education - Benchmarks First, Tools Second

PROBLEM: 1:1 tutoring can accelerate a child's learning by five months. But it is only available to those whose parents can pay, or whose schools have resources. The attainment gap - between pupils on free school meals and their peers - stubbornly persists despite decades of intervention. 470,000 teachers face rising workloads and increasing difficulty retaining staff.

THE BIG BET: This is a multi-year project. i.AI is not building an AI tutoring app. It is building the infrastructure - benchmarks, evaluation suites, data architecture - that makes safe and effective AI tutoring possible at all.

BENCHMARKS FIRST: Before any tool can be put in front of pupils, the underlying models must be proven to teach well and safely. To achieve this, robust benchmarks are being developed so that parents and teachers can be confident that AI tools for pupils are high quality, reliable and, most importantly, safe. Teachers' expertise and feedback, gathered through co-creation and school trials, will shape the datasets and standards those benchmarks rest on.

COMMON DATA ARCHITECTURE: Just putting an advanced model in front of a pupil is not enough. Learning science shows that high-quality teaching is deeply contextual. The system must know: what should this pupil be learning (national curriculum, learning objectives)? What have they already learnt (prior attainment, formative assessment, homework)? How are they being taught (lesson resources, teaching approaches)? The technical challenge is to combine this context across the classroom, anchored in the national curriculum, and make it available to the tools in a safe and privacy-preserving way

WHAT MAKES THIS DIFFERENT FROM COMMERCIAL AI TUTORING: Commercial tutoring tools are often optimised for engagement, keeping pupils on the platform. Good pedagogy sometimes requires the opposite: introducing productive struggle, correcting errors, avoiding sycophancy. The government's role is to build tools that commercial providers have no incentive to build.

TIMELINE: Eight Pioneer companies were selected to co-create tools with teachers and schools. Trials in secondary schools are scheduled for later in 2026, with impact measured through rigorous real-world trials including DfE's Edtech Testbed Programme. The DfE announcement (January 2026) committed to tools that could support up to 450,000 disadvantaged secondary pupils by the end of 2027.

Chapter 7: Scaling - From Pilot to Production

The Question the Field Keeps Avoiding

Once you have a use case that works, how do you actually get it deployed across multiple ministries? Not in theory - in operational practice. This is the question that most government AI accounts skip over, because it is harder than building the thing in the first place, and most governments have not yet solved it. i.AI has not fully solved it either - but the experience across Consult, Redbox, Minute, and Extract provides the best-documented account of what the cross-ministry scaling problem actually looks like.

The Three Structural Barriers to scaling

Infrastructure fragmentation	Different departments run on different cloud providers, different security configurations, and different data governance regimes. A tool built on Cabinet Office infrastructure cannot simply be deployed in another department - it requires months of security approvals, data linkage agreements, and infrastructure adaptation. Open-sourcing helped only where the receiving department had real engineering capability: the Department for Business and Trade stood up its own instance in a month, with hands-on support from i.AI, but most departments do not have that resource.
Ownership vacuum	Pilots exist in a political and organisational no-man's-land. No one in the receiving department fully owns the deployment decision. The central AI team built it; the department is being asked to change how they work; the champion who sponsored the pilot may have moved on by the time the security review completes. Without genuine accountability at the receiving end, scaling stalls.
Change management underinvestment	AI systems change how people work. The trust issue for Consult was not technical - it was whether civil servants and the public would trust AI-assisted analysis of democratic consultation responses. Getting to a trustworthy answer required a NESTA social readiness assessment, evaluation co-published with departments, and intentional UX friction forcing human review at every stage. None of this was retrofitted - it was baked in from the start.

Three Scaling Models - and What i.AI Actually Tried

i.AI tried three distinct approaches to scaling products across government. None is universally right. The choice depends on the product, the department, and what infrastructure already exists.

Central operation	i.AI operates the system centrally; departments use it as a service. Quality controlled, but hard to sustain at scale. Redbox experience: reached two departments in six months. Each new department required months of security approvals. Pace incompatible with ambition across 300+ organisations.
Open source self-deployment	Release code openly; departments deploy their own instances. Maximises theoretical reach but requires significant departmental engineering capability that most departments do not have. Redbox experience: self-deployment succeeded where the department had genuine engineering capability and close support from i.AI (one department stood up its own instance in a month and later diverged substantially from the original codebase), but it is not a route most departments can take, because they lack the resource to maintain a production system from source code.
Procurement-embedded deployment	Embed AI into departmental workflows through formal procurement, rather than deploying a tool and trying to integrate it afterwards. The Extract/planning approach, the MHCLG Augmented Planning Decisions programme procured under the Procurement Act 2023 and now under contract, embeds ownership at the receiving end from the start.

The Honest Scaling Heuristic: No Single Rule

The practical discipline is to identify the actual bottleneck for each product before designing the scaling approach. Consult's bottleneck is trust (hence the NESTA assessment and co-published evaluations). Extract's bottleneck is data infrastructure (hence building the data layer before any augmented decision-making). Education's bottleneck is pedagogical validity (hence the benchmark work before anything reaches a classroom). MedGuard's bottleneck was data quality (hence the pivot to research rather than deployment). Apply the solution to the actual bottleneck, not the bottleneck of the previous product.

Infrastructure as the Scaling Ceiling

The single biggest practical constraint on scaling is not model performance, funding, or political will. It is underlying data and IT infrastructure. Most English local authorities cannot deploy AI planning tools because they have not digitised their historical records. Very few clinical AI studies are conducted on real patient data, because NHS data is fragmented and inconsistently structured. Only 5–8% of clinical AI studies use real patient data, because NHS data is fragmented and inconsistently structured. The lesson: when you hit an infrastructure wall, build the infrastructure first and treat that as the product.

Chapter 8: Build, Buy, or Partner

The Framework

The build-versus-buy question is one of the most consequential decisions a government AI team makes, and the one most influenced by political economy. The practical framework i.AI has developed from experience: buy commodity AI capabilities (language models, OCR, standard classification), build custom applications of AI to specific government problems, and partner for shared infrastructure that is too expensive to build alone and too important to outsource entirely.

Situation	Recommendation	Rationale
Commodity AI (LLMs, OCR, standard classification)	Buy from established providers	No advantage to building. Market provides at lower cost and higher quality.
Custom application to a specific government problem (Consult, Extract, MedGuard)	Build in-house, with private sector components	Domain knowledge is uniquely government. AI components (model APIs, cloud) can be procured.
Infrastructure (secure cloud, data pipelines, shared APIs)	Partner or invest jointly	Too expensive for individual departments to build alone. Too important to outsource entirely.
Evaluation and audit of AI systems	Build in-house	Cannot outsource the function of checking whether vendors deliver what was promised.
Frontier AI research	Partner with academia and frontier labs	Government needs structured access to the research frontier, particularly for the Big Bet technical questions on planning, health, and education (see Chapter 2).
Bespoke advice and expertise on AI and Digital transformation	Buy from expert advisors	Bespoke, specific expertise is not efficient to develop or maintain in-house.

Using New Procurement Powers

The Procurement Act 2023 introduced the Competitive Flexible Procedure (CFP), which allows government to run iterative, dialogue-based procurement rather than the traditional specification-first approach. The MHCLG Augmented Planning Decisions tender (October 2025) is i.AI's first use of this power for an AI product - it enables government to procure AI capability without having to specify the full solution in advance, which is fundamentally incompatible with AI development.

A National AI Tender methodology, a common procurement approach that 300+ government organisations could use to prevent each from independently running procurement processes for the same capabilities, is an obvious next step.

Chapter 9: International Context - What Others Have Built

Singapore GovTech: The Most Comparable Operation at Scale

i.AI CASE STUDY | Singapore: AIBots, SENSE, and the Smart Nation Model

GovTech Singapore (~4,085 employees) manages IT for 60% of Singapore's ~100 government agencies. Annual ICT budget exceeds S\$3 billion. This is a fundamentally different scale from the UK model - but the patterns are instructive.

COMPARABLE PRODUCTS:

- VICA platform: 100+ government chatbots across 60+ agencies, 800,000+ monthly queries
- AIBots platform (launched August 2024): 40,000 users across 115 agencies in 6 months. 12,000 custom bots created. Officers can create RAG-powered chatbots in under 15 minutes
- SENSE (data assistant): converts natural language to SQL. Security model: only metadata sent to LLM providers; all data values remain within the government IT environment (Sensitive HIGH). Built by one policy officer and two GovTech engineers.

WHAT TRANSFERS: The 15-minute chatbot deployment model. The metadata-only LLM security architecture for sensitive data. The scholarship + fellowship + salary-matching talent strategy.

WHAT DOES NOT TRANSFER: Singapore's centralised system of government and unified IT estate allow faster procurement cycles. Its S\$3bn ICT budget per 5.6m population is extraordinary by any standard.

i.AI CASE STUDY | Estonia: Small Country, Replicable Architecture

Estonia's government AI rests on X-Road (open-source, MIT-licensed, 900+ organisations, 45M+ e-services/month, estimated 2% of GDP in annual savings) and Bürokratt.

BÜROKRATT: Built on RASA NLP. Core team: 12 people. Budget: €13M over four years. All code open-source on GitHub. 2025 blueprint: hybrid - Estonian Government Cloud for sensitive workloads, commercial cloud (Azure OpenAI) for agentic tasks.

WHY IT MATTERS: The X-Road model - open-source, interoperable, adopted by 25+ countries - shows how infrastructure decisions compound over decades. A 12-person team on €13M is the direct analogue to the i.AI product team model.

For Middle Powers and Developing Country Contexts: Ukraine and the Open-Source Imperative

The most instructive model for governments with limited resources and genuine urgency is not Singapore (which has extraordinary per-capita IT budgets) or Estonia (which had decades to build X-Road). It is Ukraine - a country that built one of the world's leading digital governments under full-scale war, and has now launched the world's first national AI agent for government services. The Ukrainian experience provides a practical framework.

i.AI CASE STUDY | Ukraine: Diia, Diia.AI, and the Agentic State Under War

Ukraine's digital transformation is one of the most remarkable in the world - not in spite of the full-scale Russian invasion, but partly because of the urgency it created. The starting point was near-zero digital government. Ukraine vaulted from 102nd to 5th globally in digital government services.

THE DIIA PLATFORM: Launched in 2020, Diia ('State and Me') is Ukraine's super-app for government services. More than 23 million users - over half of Ukraine's population. 30+ digital documents, 165+ services on the portal. Ukraine became the first country in the world to recognise digital passports as legally equivalent to physical ones. In the weeks following Russia's full-scale invasion in 2022, critical services were deployed in days: registration of internally displaced people, financial aid, eRecovery for rebuilding homes. 'We deployed critical services in days, not months,' wrote Ukraine's Acting Minister of Digital Transformation Oleksandr Bornyakov. Diia has been open-sourced, and exported.

DIIA.AI - THE WORLD'S FIRST NATIONAL AI AGENT: Launched in September 2025 on the Diia portal, Diia.AI is 'the world's first national AI assistant based on an agentic model - capable not only of answering questions, but actually providing government services.' Powered by Google's Gemini 2.0 Flash, fine-tuned on Ukraine's digital infrastructure. Security architecture: hybrid on-premise/cloud, with state registry data shielded from the cloud model. Currently handles 150+ services, handles more than 90% of support inquiries - human staff only step in for complex cases. The first service fully completable through chat: an income certificate. Goal: to be the personal digital assistant that proactively anticipates citizens' needs.

THE STRATEGIC CHOICE - OPEN-WEIGHT OVER PROPRIETARY: 'For our sovereign LLM, we chose Google's open-weight Gemma - because sovereignty means owning the adaptation layer, not paying for a black box.' Ukraine scores 97% on EU Open Data Maturity, third in Europe. All Diia code has been open-sourced. Strategic data sets - court documents, educational content, wartime archives - are being built as public goods.

THE DEVELOPMENT LESSON: 'Our digital state and defence innovation are not the results of top-down planning, but the fruits of open cooperation. We placed a bet on the private sector, on entrepreneurs and innovators who could move more quickly than the government.' The model was public-private partnership, open-source foundations, and political leaders who treated digital transformation as a genuine national priority - not a side project.

WHY THIS MATTERS FOR OTHERS: Ukraine's experience shatters the assumption that digital government transformation requires either peace and stability or enormous resources. The sequencing was: build the platform first (Diia), then layer AI (Diia.AI). The underlying principle was openness - open data, open source, open standards - which made adaptation fast and vendor lock-in avoidable.

The Middle Powers Argument: What the TBI Analysis Shows

The Tony Blair Institute's February 2026 paper 'Open Source: How Middle Powers Can Build Influence in the Age of AI' makes the strategic argument that applies to the vast majority of

governments - those that cannot and should not try to build frontier models, but can capture the bulk of AI's economic and strategic value through deployment, adaptation, and ecosystem building.

The core insight is that the bulk of AI's economic and strategic value will be captured downstream, through deployment, adaptation and integration into existing sectors. This is where open source becomes strategic - not as a way to build alternative national models (which quickly lag behind the frontier and become obsolete), but as a way to build the capacity to adapt, deploy, and govern AI on terms the country controls.

Don't build a national LLM from scratch	This is an inefficient use of scarce resources. National open-source models quickly fall behind the frontier and become obsolete. The value is not in the model itself but in the adaptation and deployment layer on top of it. Ukraine chose open-weight Gemma precisely for this reason.
Build the adaptation and deployment layer	The intermediary layer - fine-tuning, evaluation, inference optimisation, systems integration - is where competitive advantage sits and where governments can win. This layer compounds through shared tooling, standards, and reusable components.
Treat open-source tooling as critical infrastructure	Invest in the capacity to adapt, deploy, and govern open models. Build cost-effective public-sector tools in-house. Maintain critical open software infrastructure. This is what Estonia's X-Road did for data exchange, and what Ukraine is doing with Diia.Engine.
Build strategic datasets as public goods	High-quality national datasets are a durable source of competitive advantage. Ukraine is building court documents, educational content, and wartime archives as public goods. India's Aadhaar and UPI were data infrastructure decisions before they were AI decisions.
Use procurement to shape the open-source market	Public purchasing power should support open standards, drive startup dynamism, and build a competitive ecosystem that reduces dependence on single vendors. The UK's Procurement Act 2023 Competitive Flexible Procedure is a step in this direction.
Collaborate internationally on benchmarks	Open, sector-specific benchmarks - the UK's education benchmarks under development, Singapore's Responsible AI Playbook - are strategic infrastructure for trust and interoperability. Small countries can punch above their weight by owning the evaluation layer.

The sequencing principle holds across all contexts: digital foundations first, AI applications second. Rwanda's IremboGov (240+ services, 80% online, 5 days → 24 hours for decisions) achieved its results through platform investment before AI investment. India built Aadhaar

and UPI before the IndiaAI Mission. Ukraine built Diia before Diia.AI. The pattern is consistent: the countries that are winning on government AI built the data infrastructure first.

Chapter 10: What went wrong

A note on honesty

This chapter is uncomfortable to write. But the field of government AI desperately needs more honest accounts and fewer curated success narratives. Governments that read only the successes will try to replicate them without understanding the failures that preceded them.

- **Building before understanding users.** The most common early i.AI mistake was to build technically impressive products that did not fit how civil servants actually worked. The product manager role - embedded with end users, deeply understanding actual workflow - is more important than any technical decision,.
- **Underinvesting in scaling from the start.** The incubation function was well-designed from early on. The scaling function was an afterthought. Several technically excellent products did not reach production deployment at meaningful scale. This is the most important failure in the portfolio.
- **Not solving the infrastructure problem.** Infrastructure was always the biggest constraint on scaling and it was not solved cleanly. Building Redbox on Cabinet Office infrastructure and then trying to deploy elsewhere created enormous friction. The Extract/procurement-embedded approach is the correction, but it took too long to arrive.
- **Bureaucracy won too often.** IT access taking months, mandatory training irrelevant to the work, procurement cycles six times longer than necessary. These are individually minor and collectively devastating for team morale and productivity. Fighting this on behalf of the team is part of the leader's job - and it requires spending political capital that not every team leader has been willing to spend.
- **Not having permanent engineers for long enough.** A team that relies primarily on contract engineers cannot compound domain knowledge. Engineers who understand the codebase, the policy context, and the departmental relationships are the team's most valuable asset. Losing them at the end of a contract is not a minor inconvenience - it is a significant capability reset.

Chapter 11: The First 90 Days

Days 1–30: Foundation

- **Secure your champion.** Named minister or senior official with authority, commitment, and tenure. Without this: stop.
- **Make your first two permanent hires.** A product manager and a technical lead. These define the team's culture. Do not compromise. Do not start with contractors.
- **Address data prerequisites immediately.** If most of your target context has not digitised the relevant data (as with UK planning), the data infrastructure project is your first project - not the AI application.
- **Begin the Knowledge Hub equivalent.** Even before you have built anything, register what you are working on, what you have learned, and what others across government are doing. Prevent duplication from day one.

Days 30–60: First Project

- **Apply the Strategic Triangle.** Politically salient, technically challenging, AI-appropriate. If the project fails any one criterion, do not start it.
- **Agree evaluation criteria with stakeholders before development.** This is the Consult lesson. Criteria agreed before development are credible. Criteria designed after development to match what the model can do are not.
- **Involve security teams in the design.** Not at the deployment review - in the architecture discussion. The SENSE metadata-only model is worth examining as a default.
- **Plan the hackathon programme.** Who are the civil servant teams who will validate this product? When will you run the first hackathon? If you cannot get a room of users to a useful outcome in a half-day session, the product is not ready.

Days 60–90: Demonstrate and Learn

- **Demo real working product.** Not a presentation. A live demonstration running on real data.
- **Run the first hackathon.** The signal you need is not 'people were impressed.' It is 'teams asked to use the tool again after the session.'
- **Plan the scaling path.** Who will own this in production? What infrastructure does it need? Begin the stage gate process.

Year One Milestones

- At least one system in or near production deployment
- A clear view of the funding model for year two - whether through demonstrated savings, departmental sponsorship, or central mandate
- Permanent engineers in post across your active projects
- Team of 15–25 people with genuine technical depth

- Portfolio of 3–5 active projects at different stages of the product lifecycle
- Knowledge Hub or equivalent - at least 10 use cases registered and shared across government
- At least one published honest assessment - including what did not work

Appendix A: Decision Checklists

Project Selection Checklist

- **Political salience:** Named minister or senior official who wants this outcome and will spend political capital to unblock obstacles?
- **Technical challenge:** Interesting enough that excellent engineers and data scientists would want to work on it?
- **AI appropriateness:** High volume, pattern recognition, classification, prediction - the tasks AI actually does well?
- **Wedge vs edge:** A chokepoint in a high-volume process, or an interesting edge case?
- **Data prerequisite:** Does the data exist, is it digitised, is it detailed enough, and is using it legally permissible?
- **Evaluation framework:** Success criteria agreed with stakeholders before development begins?
- **Scaling path:** Named owner for production and clear infrastructure path?
- **Time horizon:** Committed for at least three years? If not: do not start.

Stage Gate Pre-Requirements Summary

Stage Gate	Pre-Requirements	Key Question
Scoping → Incubation	Clear problem statement. Stakeholders identified. Build vs buy assessed.	Is there a real user, clear value, and no obvious private sector solution already solving it well?
Incubation → Alpha	Demo (streamlit/notebook). Model tested. Theory of Change. Stakeholder buy-in confirmed.	Is the AI technically feasible? Is user need confirmed? Is viability plausible?
Alpha → Beta	Architecture designs. User journey proven. Reference customers identified. Sufficient model accuracy.	Does the product work? Do users find it useful enough to advocate for it?
Beta → Scaling	Reference customers achieved. Quality metrics met. Impact evaluation plan in place.	Has it been tested with a real sample in a real scenario and found to be useful?
Scaling → Graduated	Positive RCT or equivalent. Rollout plan. Handover completed. Business case secured.	Is the product self-sustaining without i.AI support?

