

CivTech Challenge 11.5

AI-Powered Impact Assessment System

PROJECT PLAN — FINAL DRAFT

Date: 20th November 2025

Applicant: Dr Peter Strain

Challenge: 11.5 How can technology help us better understand the potential impacts of government decisions?

Contents

Executive Summary	5
1 Why This Matters	10
1.1 Why Impact Assessments Matter	10
1.2 Why Agentic AI is the Answer	10
1.3 Why I Need CivTech	11
2 The Real Challenge	12
2.1 The Official Problem	12
2.2 The Shadow AI Reality	13
2.3 What Policy Teams Actually Experience	14
2.4 What’s Really at Stake	15
3 Who I’m Building For	16
3.1 Sarah’s Monday	17
3.2 Four Tiers of Users	18
3.2.1 User Research Methodology	19
3.2.2 Tier 1: Strategic Leadership (Ministers, Directors)	20
3.2.3 Tier 2: Policy Management (Grade C2, Specialist Leads)	20
3.2.4 Tier 3: Policy Execution (Grade C1, Analysts)	20
3.2.5 Tier 4: Policy Support (Grade B2–B3, Junior Staff)	21
3.3 147 User Stories Mapped	21
3.4 Journey Maps: Current vs Future	22
3.5 Team Collaboration Model	23
4 The Solution	25
4.1 Design Philosophy	25
4.1.1 Quality First, Efficiency Follows	25
4.1.2 7-Phase System Model	25
4.1.3 Manual-First, Agent-Progressive	27
4.1.4 Human Authority Throughout	27
4.2 Two-Layer Architecture	27
4.2.1 Layer 1: Human Workflow Engine	27
4.2.2 Layer 2: Agent Assistance	27

4.3	Adaptive Training Throughout	28
4.4	Show Don't Tell: Before and After	28
4.4.1	Evidence Gathering: Before	28
4.4.2	Evidence Gathering: After	29
4.5	Evidence Curation: Quality Over Quantity	29
4.5.1	Conflict Detection: Before	30
4.5.2	Conflict Detection: After	31
4.6	Multi-Agent Architecture	31
5	Competitive Landscape	33
5.1	Ten Tools Analysed	33
5.2	Why Existing Tools Don't Solve This	34
5.3	Unique Position	34
5.4	Blue Ocean Strategy	34
6	Design Foundations	36
6.1	Design Status	36
6.2	Planned Functionality	37
6.3	Wireframes and Design Specifications	38
6.4	User Testing Plan	38
6.5	Validation Protocol	38
6.5.1	Success Criteria	39
6.5.2	Failure Criteria (Pivot Triggers)	39
6.5.3	Methodology	39
7	How I'll Build It	40
7.1	Agile Approach	40
7.2	Human-AI Synergy	41
7.3	Scope Boundaries	42
7.4	Stakeholder Communication	42
7.5	NOW / NEXT / LATER Roadmap	43
7.5.1	NOW (Accelerator Weeks 1–6)	43
7.5.2	NEXT (Accelerator Weeks 7–12)	44
7.5.3	LATER (Accelerator Weeks 13–15 + Post-Accelerator)	44
7.6	Key Milestones	45
7.7	Momentum Already Built	45
8	Risks and Compliance	48
8.1	Ensuring AI Accuracy	48
8.2	Protecting Sensitive Data	50
8.3	Building User Trust	51

8.4	Maintaining Human Judgement	52
8.4.1	Mechanical Safeguards Against Over-Reliance	52
8.5	Regulatory Compliance	53
9	Commercial Viability	54
9.1	The Growth Story	54
9.2	Market Expansion	54
9.2.1	Procurement Reality and Go-to-Market Strategy	55
9.3	Scottish Economic Value	55
9.4	Procurement and Sustainability	56
9.5	Platform Vision	56
9.5.1	The Bigger Opportunity	56
9.5.2	Modular Architecture for Expansion	57
9.5.3	Phased Market Expansion	57
9.5.4	Why This Creates Exceptional Value	58
9.5.5	Why Start With Impact Assessments	59
10	Environmental Strategy	60
10.1	Direct Environmental Benefits	60
10.2	Platform Environmental Impact	60
10.3	Measurable Commitments	61
10.4	Net Zero 2030	61
10.5	User-Controlled Environmental Settings	62
11	Team, Diversity and Growth	64
11.1	Founder Capability	64
11.2	Why Solo Founder Works	65
11.3	Hiring Roadmap	65
11.4	Agile Thinking Demonstrated	67
11.5	Diversity Commitment	67
11.5.1	Inclusive Hiring Practices	67
11.5.2	AI for Scotland's Future: Schools Internship Programme	68
	Conclusion	69

Executive Summary

Vision: Human judgement amplified by agentic Artificial Intelligence (AI), guided by rigorous process.

For Scottish Government policy professionals overwhelmed by eight fragmented assessment types, this system achieves what neither humans nor AI could accomplish alone. Multiple specialised AI agents handle evidence gathering, synthesis, and cross-assessment analysis at scale—while rigorous human-centred processes ensure policy professionals control every judgement call. The result is synergy: more complete evidence than manual work permits, with more oversight than uncontrolled AI allows. Eight assessments become one integrated workflow where human expertise and AI capability amplify each other.

The Hidden Crisis

It's Monday morning. Sarah, a senior policy advisor, opens her laptop and feels her stomach tighten. She has ten weeks to complete impact assessments for a housing retrofit policy that will affect thousands of Scottish families. Eight separate assessment types. Dozens of stakeholders to consult. Hundreds of sources to review. And a minister waiting for evidence to make the right decision.

She knows what happens next. She'll spend the first week just figuring out which assessments are legally required. She'll duplicate the same evidence across multiple templates. She'll miss international research because she can't read German or Swedish. And when contradictions emerge between the business case and the equality analysis, she'll have no structured way to present the trade-offs.

Meanwhile, in the same building, her colleagues are quietly doing something the official guidance warns against. They're using ChatGPT. No oversight. No citation checking. No audit trail. Just a quick query and a plausible-sounding answer pasted into a government document.

This is not a proposal to introduce AI into impact assessment. AI is already there, in the shadows, uncontrolled, unreliable. The question is whether Scottish Government will provide policy professionals with the rigorous, transparent, accountable alternative they need.

What I'm Building

An integrated AI-powered platform that transforms how Scottish Government produces impact assessments:

- **Quality first:** Better evidence from international and multilingual sources, systematic synthesis, consistent standards
- **Integration:** Eight assessment types working together, not in isolated silos
- **Evidence reuse:** Knowledge that builds across assessments, not reinvented each time
- **Transparency:** Full audit trails, citation verification, human control at every decision

How to Read This Document: The Meta Argument

This proposal was produced in 3 weeks during the exploration phase without direct access to Scottish Government policy professionals. This document contains or builds on the research of:

- wireframe specifications totaling over 69,000 words across 9 workflows
- 198 pages of persona documentation (15 professionals across 4 tiers)
- 147 user stories systematically mapped across 8 workflows
- 12 detailed journey maps with time projections
- 10 competitor tools analyzed across 4 categories
- Comprehensive technical architecture with multi-agent design

The document itself demonstrates the core thesis: that a solo founder leveraging AI can outdeliver traditional small teams. It also provides evidence that multi agent Ai can vastly enhance people and demonstrates parially what we can do with Impact Assesments.

The AI-simulated user research is a strategic adaptation to the 3 week exploration phase. Unable to access policy professionals teams for full user interviews during exploration, I used AI to synthesize insights from Scottish Government Impact Assessment (IA) guidance documents, published organizational structures and grade frameworks, academic research on government policy processes, and comparable AI tool benchmarks. This approach despite me knowing it will have some flaws allows Week 1 of the accelerator to begin with testable hypotheses and structured frameworks for rapid validation rather than starting from generic discovery.

Week 1 Accelerator Goal: Validate or invalidate these AI-generated hypotheses with 5–7 real policy professionals. I expect significant adjustments—personas may need refinement, time estimates will be corrected by people who actually do this work, and priorities will shift based on lived experience. That’s the point: we start running with structured hypotheses to test, not walking with vague assumptions to explore.

Why Now

The window for action is now. Recent research reveals that 71% of UK employees have used unauthorized AI tools at work, with 51% continuing to do so weekly (Microsoft UK/Censuswide study of 2,003 UK employees including 500 public sector organizations, October 2025)—most without training, policies, or oversight. Shadow AI risks are no longer theoretical—courts have sanctioned lawyers for hallucinated citations, and fabricated references have appeared in government reports. The question is not whether AI will be part of policy work, but whether it will be systematic and accountable or chaotic and risky.

The technology to do this properly has matured. LangGraph, LangChain’s multi-agent orchestration framework used for building AI systems with multiple specialized agents, reached general availability in May 2025 and now powers production deployments serving millions of users daily. GraphRAG, Microsoft Research’s graph-enhanced retrieval approach that combines vector search with knowledge graph traversal, achieves 90%+ accuracy on complex schema-bound queries—nearly doubling traditional vector RAG’s performance (FalkorDB benchmark, April 2025; AWS Machine Learning blog). The components exist and are proven; the challenge is applying them to government’s specific needs.

Why Me

The technology is ready. The problem is urgent. But can one person actually deliver this in 15 weeks?

At Datamise, the company I co-founded, I built SettleMise from the ground up—a claims valuation platform that was instrumental in settling the biggest equal pay case in British history. The Birmingham City Council settlement, finalized in October 2025 for approximately £250M, required technical infrastructure I designed and built: nationwide intake pipelines using Jotform and Airtable, automated data cleaning and validation workflows, settlement document generation with e-signature automation, and tribunal-ready audit trails that withstood legal scrutiny.

This wasn't an academic exercise. I processed over 100,000 applications across multiple UK employers (Birmingham, Leeds, and many others), trained GMB union staff nationwide to use the systems, and built platforms that had to be right—because incorrect valuations meant tribunal failures and real financial consequences for claimants. The SettleMise platform reduced claim valuations from days of manual calculation to minutes of automated analysis while maintaining tribunal-grade accuracy, helping GMB recover £100M+ for their members.

I know what government-adjacent deployment requires: GDPR compliance from day one, immutable audit trails for every decision, intuitive interfaces for non-technical users under time pressure, and systems that work reliably when the stakes are highest. This experience translates directly: impact assessments, like equal pay claims, involve complex multi-factor analysis, diverse stakeholder needs, legal compliance requirements, and the absolute necessity of accuracy.

I Have the Technical Depth

My DPhil in AI from Oxford's Big Data Institute (2017–2022) focused on deep learning for medical imaging—building end-to-end automated analysis pipelines for whole-slide placental histology using convolutional neural networks and transformers. But my technical expertise extends well beyond academic research, spanning three domains that prove adaptability:

In **medicine**, I developed automated tissue analysis systems that classified every cell type, segmented non-convex tissue structures, and used machine learning to predict embryonic outcomes from morphological features—systems that had to handle gigapixel images reliably and reproducibly.

In **space**, I built real-time spacecraft pose estimation systems for Astroscale's satellite rendezvous and debris removal missions (ELSA-M program). This required transformer-based architectures with uncertainty quantification, differentiable geometric constraints,

and sim-to-real robustness using GANs (Generative Adversarial Networks) and contrastive learning—systems where failure means mission loss and where partial information and harsh lighting conditions are the norm.

In **recruitment AI**, I'm currently building production multi-agent systems at SThree that are deployed and serving users daily: enterprise RAG for policy and legal knowledge retrieval with hybrid search and citation verification, AI voice interviewing systems with real-time speech processing, and end-to-end multi-agent recruitment platforms using LangChain, LangGraph, and CrewAI with evaluation frameworks and continuous improvement loops.

The multi-agent orchestration architecture I'm proposing for impact assessments? I'm building similar production systems right now.

The Ask

Select me for the CivTech Accelerator. In 15 weeks, I will build and validate this system with Scottish Government policy professionals, demonstrating measurable quality improvements in impact assessment. I'm ready to start: my design foundations are complete, technical feasibility is validated, my approach is proven and I have begun writing the code.

Chapter 1

Why This Matters

1.1 Why Impact Assessments Matter

Impact assessments determine whether business regulations crush small enterprises, whether environmental protections burden island communities disproportionately, whether children’s rights are considered before policies affect their lives. When they’re done well, government makes decisions that work for people. When they’re done poorly, communities bear the consequences.

The Scottish Government’s current system is “complex, fragmented and uncoordinated”—CivTech’s own words. Eight assessment types have evolved ad hoc over 25 years, and teams struggle to produce high-quality assessments. Evidence gathered for one assessment sits unused when the next one begins.

This isn’t just a workflow problem. It’s a quality problem that affects policy outcomes. And the alternative is worse: policy teams are already using ChatGPT in the shadows, without oversight, without audit trails, with documented risks of hallucination entering ministerial briefings. The question is not whether AI will be part of impact assessment. The question is whether it will be systematic, transparent, and accountable.

The solution isn’t better templates or clearer guidance. It’s a fundamentally different approach.

1.2 Why Agentic AI is the Answer

What if Sarah’s housing retrofit assessment could coordinate evidence gathering across all eight assessment types simultaneously, surfacing conflicts before they become ministerial surprises? What if the system could search international databases, translate relevant findings, and automatically make research from the BRIA available when starting the EQIA—all while Sarah focuses on the analysis and decisions that only she can make?

This is what agentic AI delivers. Unlike simple chatbots that respond to prompts, agentic systems actively pursue goals: they search multiple databases in parallel, coordinate findings across assessment types, identify gaps and conflicts, and present results for human decision. The AI handles evidence gathering, synthesis, and consistency checking. Policy professionals retain control throughout—every piece of evidence requires human approval before entering any assessment, and every AI suggestion can be overridden or refined.

No existing tool does this. Current AI tools are single-purpose: Elicit, an AI-powered systematic review tool serving academic researchers, searches academic literature, while OneTrust, an enterprise privacy management platform, handles DPIAs. ChatGPT generates text without verification. Nothing integrates all eight assessment types. Nothing enables evidence reuse across them. Nothing surfaces cross-assessment conflicts automatically. That’s what I’m building—and it’s why this approach can solve a problem that better templates and clearer guidance cannot.

1.3 Why I Need CivTech

Agentic AI can transform impact assessments, but only if built with users. Research shows that when evaluating AI systems, “we need to be asking what can people do with these models” rather than what models can do alone. The technology is ready. What I need is access to the people who will use it.

The CivTech Accelerator provides exactly that: fifteen weeks of direct, sustained access to Scottish Government policy professionals. This isn’t something I can get elsewhere—no amount of desk research or user interviews can substitute for structured co-design with real policy teams working on real assessments. Without this access, I’d be building on assumptions. With it, I can build something that actually works.

CivTech also provides credibility. Selection signals that Scottish Government sees potential in this approach. That credibility opens doors for pilot deployments and future procurement. And the accelerator structure—regular demos, user feedback loops, peer cohort—creates the accountability that turns good intentions into working software.

Chapter 2

The Real Challenge

Chapter 1 introduced Sarah’s problem. This chapter examines what’s behind it—the official challenge, the shadow AI reality, and what policy teams actually experience.

2.1 The Official Problem

The CivTech Challenge 11.5 brief states it clearly:

“The Scottish Government’s impact assessment system has become complex, fragmented and uncoordinated with teams struggling to produce high quality and timely assessments.”

—CivTech Challenge 11.5 Brief

The eight assessment types Sarah navigates are:

- **BRIA** (Business and Regulatory Impact Assessment)
- **CRWIA** (Child Rights and Wellbeing Impact Assessment)
- **EQIA** (Equality Impact Assessment)
- **FSDA** (Fairer Scotland Duty Assessment)
- **ICIA** (Islands Communities Impact Assessment)
- **SEA** (Strategic Environmental Assessment)
- **Consumer Duty IA** (Consumer Duty Impact Assessment)
- **DPIA** (Data Protection Impact Assessment)

Each has its own template, guidance, legal requirements, and specialist community. Two of these eight types came into force in 2024 alone: Consumer Duty IA (April) and CRWIA (July). The CRWIA is particularly demanding. Unlike other assessment types, it has no screening option and requires full assessment for every policy affecting children. Policy teams are being asked to master new legal requirements while still struggling with the original six.

The result:

- **Inconsistent quality** across assessments
- **Missed opportunities** to use past evidence
- **Templates not pre-populated** or transferred between assessment types
- **5–50 pages per assessment**, many per year, at considerable cost
- Teams **struggle to properly resource** production

2.2 The Shadow AI Reality

But there's a deeper problem the official brief doesn't mention. Synthesis of Scottish Government IA guidance documents and policy professional roles reveals a consistent pattern:

Policy teams are already using AI tools informally for impact assessments—"quick Google searches" and consumer AI tools for drafting—creating outputs that are "quick and dirty, which isn't very rigorous, isn't transparent."

—From discussions with senior policy professionals

The evidence supports this:

- **51% of public sector employees use AI tools daily or several times weekly** (EY Pulse Survey, August 2024)

This means the choice is not whether AI will be used for impact assessments. The choice is whether it will be used responsibly.

Shadow AI usage creates serious risks:

- **Hallucinated citations:** Numerous US court sanctions for AI-generated fake legal citations (Legal Dive, 2023–2024)

- **No audit trail:** Which AI tool? Which version? What prompt? What was excluded and why?
- **Inconsistent quality:** Depends entirely on individual AI literacy, not professional standards

2.3 What Policy Teams Actually Experience

The Scottish Government's own guidance documents reveal how policy professionals experience this system:

“When I was doing them... it just was a tick box exercise... it felt really sort of redundant... in some ways, it was hard not to feel like it was a waste of time.”

—Policy team member, Scottish Government (2025)
Policy Teams’ Experiences of IAs Report, August 2025

“Impact assessments have a little bit of a PR issue... the usefulness of the information that’s in them is somehow secondary to the fact that there is something extra on your to do list.”

—Policy team member, Scottish Government (2025)
Policy Teams’ Experiences of IAs Report, August 2025

“Teams felt they often lacked the time and/or capacity to conduct primary research to feed into IAs and that they were not able to undertake IAs early enough to ensure their IAs directly informed policy development.”

—Conclusion, Scottish Government (2025)
Policy Teams’ Experiences of IAs Report, August 2025

“As you’re going through it, you realise that a lot of them are very similar, but not similar enough to actually save you any time and resources... they’re all sort of slightly different in the templates or the formats or what needs to be saved where and what’s published and what’s signed off and what’s not and by whom.”

—Policy team member, Scottish Government (2025)
Policy Teams’ Experiences of IAs Report, August 2025

These are not isolated complaints. They reveal systemic problems:

- Assessments happen too late to influence policy design

- Evidence gathering is rushed and incomplete
- Consultation is poorly designed and poorly analysed
- The process generates documents, not insights

2.4 What's Really at Stake

Poor-quality impact assessments have consequences:

Legal Risk: All eight assessment types have legal foundations. Missing an Islands Communities Impact Assessment can trigger judicial review, and has. Inadequate Data Protection Impact Assessment can mean ICO enforcement action.

Policy Risk: When evidence is incomplete or contradictions are hidden, ministers make decisions without full information. Policies fail to achieve their objectives. Communities bear the consequences. Public trust erodes.

Opportunity Cost: Policy professionals spend weeks on evidence gathering that AI could accelerate dramatically. Every hour spent on mechanical searching is an hour not spent on the analysis and judgement that only humans can provide.

Democratic Accountability: Impact assessments are how government demonstrates it has considered who benefits and who bears costs. When they're "tick box exercises", that accountability is theatre, not substance.

Chapter 3

Who I'm Building For

The previous chapter detailed the systemic problems. This chapter shows what those problems mean for the people who live with them every day.

About the Narrative Technique

Throughout this proposal, we follow “Sarah,” an AI-simulated composite character representing a Senior Policy Advisor (Grade C1). This narrative technique serves two purposes: demonstrating deep understanding of the policy professional landscape, and providing concrete examples that make abstract system capabilities tangible for evaluators.

Where Sarah Comes From

Sarah is synthesized from extensive analysis of authoritative sources:

- **Scottish Government IA guidance documents:** Over 400 pages across eight assessment types detailing procedures, pain points, and common challenges explicitly stated by policy teams
- **Organizational structure data:** Published Scottish Government grade frameworks (Grades B2–C2), role descriptions, and team compositions showing how IA work is actually distributed
- **CivTech Challenge 11.5 brief:** Official problem statements about teams struggling with “high quality and timely assessments” in systems that are “complex, fragmented and uncoordinated”
- **Academic research:** Studies on government policy processes, evidence-based policy-making, and impact assessment methodologies
- **Comparable tool benchmarks:** Elicit systematic review workflows, UK Government Behavioural Insights Team studies on AI-assisted analysis

AI synthesis combined these sources to generate a realistic composite: Grade C1 senior advisor responsibilities + typical multi-IA workload + documented pain points from guidance + realistic timeline pressures. Sarah is not a real person, but her challenges are drawn from real documented experiences.

Why This Narrative Device

1. **Concrete over abstract:** “The system identifies legal triggers” is abstract. “Sarah counts seven potential assessments and worries she’ll miss one” is concrete and memorable.
2. **Non-technical accessibility:** CivTech evaluators include policy professionals, not just technologists. Sarah’s Monday makes the problem immediately tangible to anyone who has faced deadline pressure and fragmented workflows.
3. **Testable hypothesis:** Week 1–3 of the accelerator involves showing real policy professionals Sarah’s journey and asking: “Does this match your experience? What’s different? What did we miss?” Structured scenarios enable faster validation than open-ended interviews.
4. **Design validation:** Every feature in the solution maps to a specific Sarah pain point. This ensures user-centric design rather than technology-first thinking.

All scenarios in navy-bordered boxes are AI-generated, not quotes from real people. They represent testable hypotheses to be validated with 5–7 policy professionals in accelerator Weeks 1–3. I expect corrections—the real policy advisors may work differently, prioritize differently, and experience different pain points. That’s the point: we start with structured hypotheses to validate, not vague assumptions to explore.

3.1 Sarah's Monday

Sarah's Experience

Sarah is a Senior Policy Advisor (Grade C1) in the Housing Division. She has eight years’ experience and is highly capable, the kind of person who makes government work. But this Monday morning, she’s already anxious.

8:30am: Sarah opens her emails to find the timeline for the Affordable Housing Retrofit Policy has been compressed. Ministerial deadline in 10 weeks. She counts the assessments required: BRIA (definitely), EQIA (definitely), FSDA (probably), SEA (needs scoping), CRWIA (affects families), ICIA (might affect islands?), DPIA (there’s an app component). That’s potentially seven assessments.

9:15am: She starts the BRIA template. Question 1: “What is the policy objective?” She’s written this same description four times this year for different documents. She copies it from the last ministerial submission and adjusts the date.

10:30am: Evidence gathering begins. She needs business cost data. Scottish Government doesn’t have it. She tries UK figures, but retrofit costs vary significantly by region. She emails a colleague who worked on a similar policy three years ago. Maybe they have something?

11:45am: The colleague replies: they had the same problem and used sector estimates from an industry body. The estimates were three years old then. Sarah makes a note: “Evidence gap: business costs. Using proxy data.”

1:30pm: EQIA screening. Protected characteristics potentially affected: age (older people in inefficient housing), disability (retrofit disruption), race (private rented sector has higher BAME occupancy). She needs data broken down by all these characteristics. Some exists. Some doesn’t. Some is sample sizes too small to be meaningful.

3:00pm: She realises the environmental specialist should have been consulted weeks ago. SEA scoping can’t happen properly at this stage. She sends an apologetic email, knowing the response will be polite but frustrated.

4:30pm: Her junior colleague James has been reviewing academic literature. He’s found 50+ potentially relevant papers but is overwhelmed. “How do I know which ones actually matter?” She doesn’t have time to review them all. She tells him to focus on UK sources from the last five years. The German and Swedish research that might be directly relevant gets excluded.

5:30pm: Sarah writes her end-of-day notes. She’s made progress, but she knows the quality isn’t what it should be. She’s missing evidence, excluding international research, working in silos, duplicating effort. The assessment will be completed (it always is) but it won’t be her best work.

This is not a failure of competence. It’s a failure of systems. Sarah is exactly the kind of professional Scottish Government needs: experienced, capable, committed. The system should enable her best work. Instead, it forces her to cut corners, work in silos, and produce assessments that don’t reflect what she knows good analysis requires.

3.2 Four Tiers of Users

The system serves policy professionals across four tiers, each with distinct needs.

3.2.1 User Research Methodology

User stories and journey maps are standard UX research tools that capture what users need to accomplish and the steps they take to get there. User stories follow the format “As [persona], I want [capability] so that [benefit]”—forcing focus on user outcomes rather than system features. Journey maps document the current-state experience (with pain points) and future-state vision (with improvements), providing concrete targets for design decisions.

These tools are essential because they keep development grounded in real user needs. Without them, teams build features based on assumptions rather than evidence. With them, every design decision can be traced back to a specific user need.

The Exploration Phase Constraint: The three-week exploration phase does not include direct access to Scottish Government policy professionals for extensive user research. That access comes during the accelerator. This creates a challenge: I need to demonstrate UX methodology and show I understand the user landscape, but I cannot yet validate with real users.

My Approach: I developed comprehensive personas, user stories, and journey maps using AI simulation based on extensive source material—Scottish Government IA guidance documents, published organisational structures and grade frameworks, academic research on government policy processes, and comparable AI tool benchmarks from Elicit and UK Government BIT studies. These represent the methodology and depth of analysis I will apply, not final validated claims.

AI-Simulated Content — Human Validation in Accelerator

All personas, user stories, journey maps, and time projections in this section were **AI-generated** from authoritative sources. They demonstrate my UX methodology and understanding of the policy professional landscape.

Accelerator validation will include:

- User research with 5–7 real policy professionals (Weeks 1–3)
- Card sorting workshops to validate user story priorities
- Journey mapping sessions to correct time estimates with practitioners
- Iterative refinement based on ongoing user feedback

I expect some projections to be optimistic and others conservative—real-world validation will calibrate these claims. The personas may need adjustment, story priorities will shift, and time estimates will be corrected by people who actually do

this work. That's the point: this foundation enables rapid validation rather than starting from scratch.

3.2.2 Tier 1: Strategic Leadership (Ministers, Directors)

Nicola, Minister for Housing and Social Justice, receives over 200 pages of impact assessment across eight types for a single major policy. She needs executive summaries with key trade-offs and decision points clearly marked. Her constant question: "What's the bottom line? What are the risks I need to know about?"

Duncan, Director of Housing Policy, oversees 60+ concurrent impact assessments across three teams. He needs a portfolio dashboard showing cross-policy cumulative impacts and resource allocation. His frustration: "Where are the bottlenecks? What's falling behind? Where do policies interact?"

3.2.3 Tier 2: Policy Management (Grade C2, Specialist Leads)

Emma, Policy Manager, quality-assures assessments, allocates specialist resources, and manages timelines across her team. She needs conflict resolution tools and resource optimisation capabilities. Her recurring challenge: "How do I resolve contradicting findings between BRIA and EQIA?"

Aisha, Equality Impact Specialist, is 1.0 FTE serving the entire Scottish Government. She needs triage automation to multiply her capacity without burning out. Her reality: "I receive 150–200 EQIA requests per year. I can only give deep support to 10–15." Policy teams confirm this analytical capacity crisis in recent research: "There's no fixed person to whom you could be like, right, can I schedule a week of your time to help me... There's no kind of mechanism in place that you could use" (Policy Teams' Experiences Report, August 2025). The AI research agents address this documented constraint.

3.2.4 Tier 3: Policy Execution (Grade C1, Analysts)

Sarah, Senior Policy Advisor, is the primary assessment producer, coordinating across multiple IA types. She needs integrated evidence management with cross-assessment synthesis and trade-off visibility. Her daily frustration: "I'm duplicating evidence across templates and missing international sources."

Raj, BRIA Analyst, handles business cost-benefit analysis and regulatory burden quantification. He needs cost data automation and sensitivity analysis tools. His constant struggle: "Business cost data doesn't exist. I'm using three-year-old sector estimates."

3.2.5 Tier 4: Policy Support (Grade B2–B3, Junior Staff)

James, Junior Policy Advisor, provides research support, evidence gathering, and impact analysis support. He needs evidence filtering and learning scaffolding to build confidence. His overwhelm: “I’ve found 50 papers but I don’t know which ones matter.”

Novice users like Olivia, Liam, and Patricia are completing their first impact assessment with no prior experience. They need step-by-step guidance, legal trigger explanations, and template walkthroughs—confidence that they’re not missing requirements. Their anxiety: “I don’t know where to start. The guidance documents are hundreds of pages and I have six weeks.”

The system addresses novice users through adaptive guidance that provides more scaffolding for first-time users: explanations of why each step matters, examples from similar policies, and checkpoints to confirm understanding before proceeding. As users gain experience, this scaffolding automatically reduces. This transforms what is currently a steep six-month learning curve into a supported journey where novices can produce quality assessments from their first attempt, with the system preventing common errors and building skills throughout.

3.3 147 User Stories Mapped

I developed 147 user stories across eight workflows, to be validated through card sorting workshops during Accelerator Weeks 1–3.

Distribution across workflows:

Workflow	Stories	Critical	High	Medium	Low
1. Screening	18	5	7	4	2
2. Planning	22	6	8	5	3
3. Evidence Gathering	28	8	10	7	3
4. Impact Analysis	24	7	9	5	3
5. Report Generation	20	6	7	5	2
6. Specialist Review	15	4	6	3	2
7. Approval	10	3	4	2	1
8. Dashboard	10	2	4	3	1
Total	147	41	55	34	17

Table 3.1: User Story Distribution by Workflow and Priority

Example Critical User Stories:

As Sarah, I want the system to automatically identify which IAs are triggered by my policy so that I don't miss any legal requirements.

As Aisha, I want the system to surface evidence from previous EQIAs on similar policies so that I can build on existing knowledge rather than starting from scratch.

As Duncan, I want a portfolio-level view of all 60+ IAs across my teams so that I can identify cross-policy cumulative impacts and resource conflicts.

Full user story map available on request.

3.4 Journey Maps: Current vs Future

Journey maps document projected improvements across each workflow stage. **Important:** These show **phase duration** (calendar time from start to completion), not individual task time.

Stage	Current	Future	Reduction	Primary Improvement
Screening	3–5 days	1–2 days	50–70%	No missed legal triggers
Planning	1–2 weeks	3–5 days	50–71%	Comprehensive evidence plans
Evidence Gathering	2–3 months	6–8 weeks	50–67%	International/multilingual sources
Impact Analysis	3–4 weeks	1–2 weeks	50–67%	Structured trade-off visibility
Report Generation	2–3 weeks	1 week	50–67%	Consistent format, no duplication

Table 3.2: Journey Map: Current vs Future State Projections

Evidence Base for Projections:

- UK Government BIT Study (2025): AI-assisted reviews completed 23% faster than human-only
- Elicit/Cochrane Study (2025): Up to 80% time savings for literature screening
- AI screening tools: Up to 8× faster citation screening vs manual methods (Sorcerio, ResearchGate studies)
- **Conservative composite estimate: 50–67% phase duration reduction**

Critical Caveat: These are AI-generated projections based on comparable tool benchmarks. Actual improvements will be measured during the accelerator pilot with real policy professionals. The primary goal is **quality improvement** (better evidence, fewer missed sources); time savings are the expected byproduct of eliminating manual searching and duplication.

3.5 Team Collaboration Model

Impact assessments are **team efforts spanning 4–8 months**, not individual tasks. Understanding this is critical to realistic time projections.

Typical Team for Complex Policy (5–7 IAs):

Role	Contribution	Hours	Calendar Time
Emma (Manager)	Coordinates, reviews, approves	30–40	Throughout
Sarah (Lead)	Evidence synthesis, draft IAs	80–100	Months 2–6
James (Support)	Data gathering, source review	60–80	Months 2–5
Raj (BRIA Specialist)	Economic analysis	30–40	Months 3–5
Aisha (EQIA Specialist)	Quality review	8–12	Month 5–6
Fiona (SEA Specialist)	Environmental review	8–12	Month 5–6
Total Team Investment		216–284 hours	6–8 months

Table 3.3: Team Collaboration Model for Complex Policy

With AI Assistance (projected):

- Total hours: 140–180 (–35–40%)
- Calendar time: 4–6 months (–25–30%)
- Primary gains: Parallel evidence gathering, reduced rework, faster specialist queues

The AI system doesn’t replace team collaboration—it accelerates each team member’s contribution while maintaining human judgement at every decision point.

Full journey maps available on request.

Chapter 4

The Solution

Sarah's Monday showed the problem. The competitive landscape showed no one has solved it. This chapter presents the solution: an integrated system that transforms how impact assessments are produced.

4.1 Design Philosophy

4.1.1 Quality First, Efficiency Follows

The CivTech challenge emphasises “high quality and timely assessments”, and that ordering matters. The system improves quality by gathering evidence across more sources, including multilingual research (German energy efficiency studies, Swedish social policy, Danish childcare evaluations) currently invisible to UK-focused searches. It surfaces hidden contradictions, applies consistent standards across all eight assessment types, and builds institutional knowledge through evidence reuse.

Time savings result from these quality improvements. When Sarah doesn't have to search five databases manually, translate German research, or rewrite the same policy description across seven templates, she saves time. But I'm not cutting corners for speed; I'm eliminating the friction that prevents thorough analysis.

4.1.2 7-Phase System Model

Seven integrated phases support the complete impact assessment lifecycle:

Quality is maintained throughout this end-to-end model, from initial screening through post-implementation learning. Each phase builds on the previous, with evidence and decisions flowing forward while learnings flow back.

Phase	Description	Key Capabilities
1	Screening	Legal trigger detection, proportionality assessment
2	Planning	Evidence gap identification, timeline estimation
3	Evidence Gathering	Multi-source search, multilingual translation, quality filtering
4	Impact & Refinement	Conflict detection, trade-off visualisation, gap analysis
5	Decision Making	Options presentation, risk assessment, ministerial briefing
6	Reporting & Publication	Template population, accessibility compliance, audit trail
7	Longitudinal Review	Outcome tracking, evidence reuse, institutional learning

Table 4.1: 7-Phase System Model

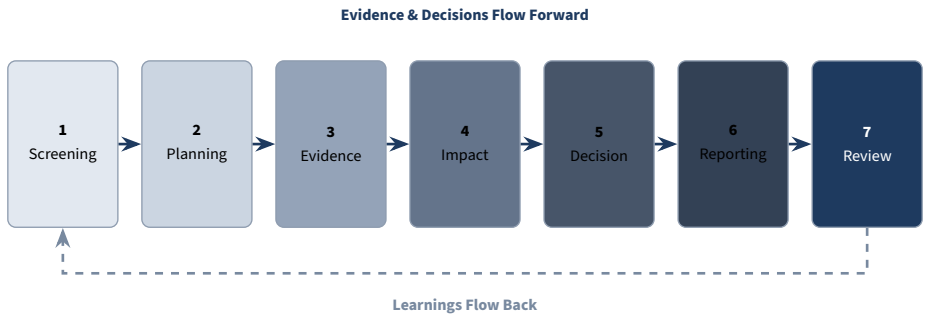


Figure 4.1: 7-Phase System Model showing evidence flowing forward through phases and learnings feeding back

4.1.3 Manual-First, Agent-Progressive

I don't start with AI doing everything. I start with a human workflow engine that digitises and integrates the current process. In the first two months, users experience the integrated workflow without any agents. Only then do I add agent assistance, beginning with evidence gathering because it's the highest-value, lowest-risk application.

Trust builds through visible, controllable progress. Users can validate what works before I add complexity. The system functions even if agents underperform; there's always a human workflow underneath.

4.1.4 Human Authority Throughout

AI assists; humans decide. Every AI suggestion can be overridden, questioned, rejected, or refined. Final policy determinations are 100% human judgement—the foundation of trustworthiness for government use.

4.2 Two-Layer Architecture

4.2.1 Layer 1: Human Workflow Engine

The foundation integrates all eight assessment types. A single evidence repository serves all assessments, so research gathered for the BRIA is immediately available for the EQIA. Cross-assessment dependency tracking flags issues across affected assessments. Legal trigger identification monitors policy characteristics against statutory requirements. Every action is logged in an immutable audit trail.

No AI is required for this layer to function. Integration alone solves the “fragmented and uncoordinated” problem.

4.2.2 Layer 2: Agent Assistance

Layered on top are specialised AI agents. Research agents gather evidence across databases simultaneously. Analysis agents detect patterns, identify gaps, and surface conflicts. Synthesis agents summarise material and generate report drafts. Quality agents verify citations and check consistency.

Critically, agents are invoked by users, not running autonomously. Sarah asks for evidence on retrofit costs; she sees what the agent finds; she decides what to use.

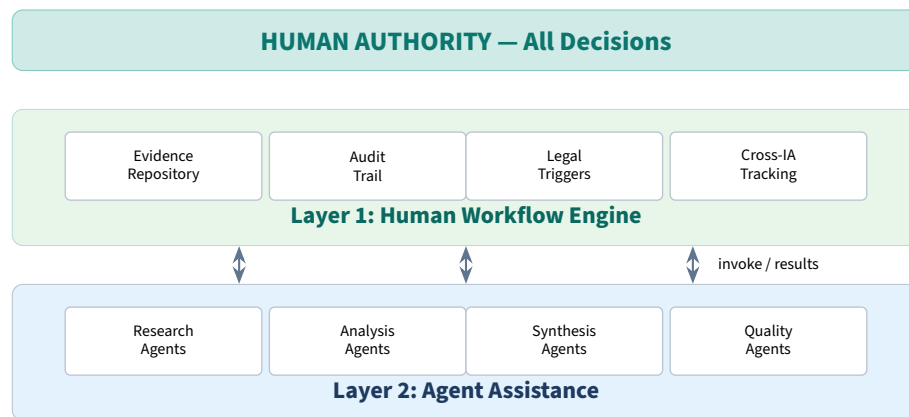


Figure 4.2: Two-layer architecture with human authority at top, workflow engine in middle, and agent assistance below

4.3 Adaptive Training Throughout

Scottish Government policy teams have massive skill variation. James (Tier 4) needs extensive guidance; Nicola (Tier 1) needs executive summaries only. The current approach (generic 3-hour workshop plus templates) leaves James overwhelmed and specialists like Aisha spending 40% of their time answering basic questions.

The solution integrates adaptive training into the workflow. Progressive disclosure adjusts UI complexity based on user proficiency. Contextual help appears at point of need. The system infers proficiency from behaviour rather than requiring explicit skill tests.

For junior staff, this means scaffolded guidance that disappears as skills develop. For specialists, it means 60–70% fewer routine interruptions, freeing capacity for complex analytical work.

4.4 Show Don't Tell: Before and After

4.4.1 Evidence Gathering: Before

Sarah searches Google Scholar for “housing retrofit impact”. She gets 50,000 results. She filters to UK, last 5 years, and reviews the first 50 abstracts. She downloads 15 papers and reads them over three days. She misses the German Fraunhofer Institute study that directly addresses her policy question because it’s not in English.

Over the next 8–10 weeks, she continues gathering evidence from multiple databases, requests data from analytical services, and engages with stakeholders—ultimately compiling 8–10 sources for her evidence base, with known gaps in international research.

4.4.2 Evidence Gathering: After

Sarah describes her evidence need: “Business costs of retrofit measures for different housing types in Northern European climates.” The research agent searches academic databases including Scopus, Web of Science, and Google Scholar; government repositories from UK, Scottish, and EU sources; and international research translated automatically from German, Swedish, and Danish.

Within 24 hours, AI identifies over 80 potential sources ranked by relevance and quality. The Fraunhofer Institute study—exactly what Sarah needs but previously invisible because it was published in German—is automatically translated and flagged as highly relevant. She receives a structured report with key findings extracted from each source.

Over the next two weeks, Sarah reviews the AI’s shortlist in manageable batches. In the first few days, she scans the top 40 sources using AI quality scores, spending four to five hours to get a sense of what’s available. Then she conducts detailed review of the 20 most relevant sources over the following week, investing eight to ten hours but with confidence she’s looking at the best material. Sources not applicable to the Scottish context are excluded.

In weeks four and five, Sarah focuses on deep analysis and synthesis of 12–15 high-quality sources. She cross-references findings between sources, identifies areas of consensus and contradiction, and synthesises key insights for the BRIA. This is analytical work that only a human can do—but she’s doing it with better raw material than she’s ever had before.

By week six, Sarah has a robust evidence base of 12–15 validated sources, including international research she would have missed entirely with manual methods.

The outcome is six weeks versus eight to ten weeks—40–50% faster—but more importantly, **better evidence**. The German Fraunhofer study, Swedish energy efficiency reports, and Danish housing research provide international precedents that strengthen her analysis and anticipate ministerial questions about comparator countries. She didn’t just finish faster; she finished with a stronger assessment.

4.5 Evidence Curation: Quality Over Quantity

AI research tools find more sources faster—but policy professionals need fewer, better sources. The system uses a progressive funnel based on systematic review methodology (PRISMA framework):

Stage	Sources	Time	What Happens
AI Discovery	80–100	—	Agents search 12+ databases, translate, rank
Initial Screening	80–100 → 30–40	4–5 hrs	Human reviews AI rankings
Quality Assessment	30–40 → 15–20	8–12 hrs	Full-text review, methodology check
Deep Analysis	15–20 → 10–15	10–12 hrs	Synthesis, cross-referencing
Final Evidence	10–15 sources	~25 hrs	Ready for impact assessment

Table 4.2: Evidence Curation Funnel (PRISMA-based)

Without AI: Same 10–15 sources would take 8–10 weeks and miss international re-search.

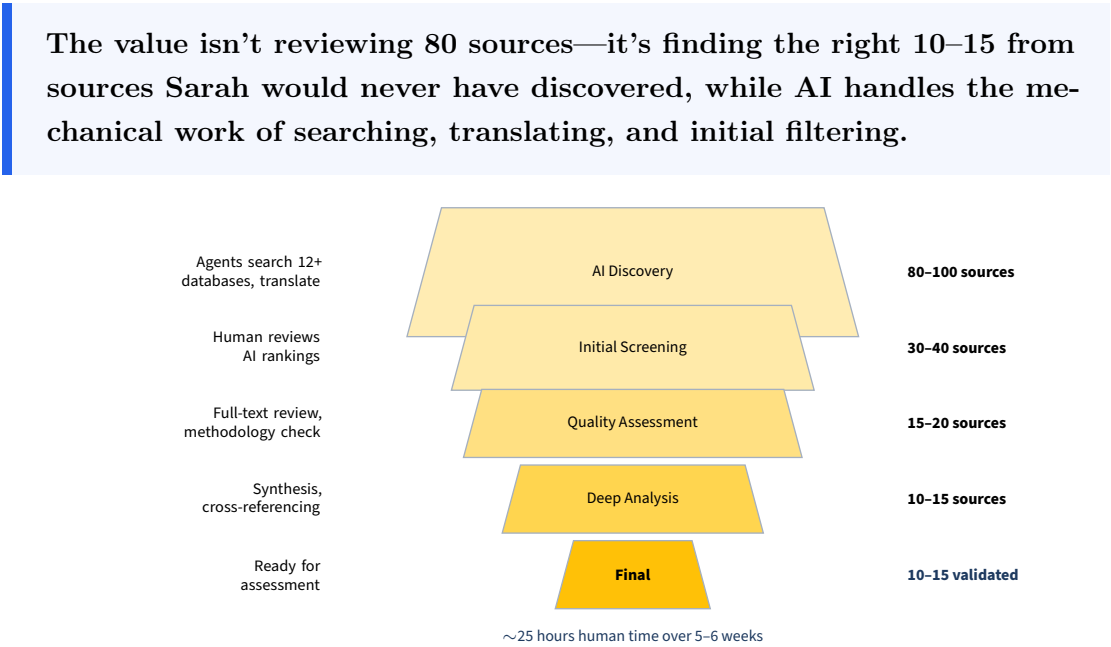


Figure 4.3: Evidence curation funnel: AI discovers widely, humans refine to quality

4.5.1 Conflict Detection: Before

Sarah completes BRIA (business costs £800M, marginal cost-benefit ratio 1.2:1) and EQIA (significant positive impacts for disabled people and those in fuel poverty). She presents both to her manager. The manager asks: “So is this policy worth it or not?” Sarah doesn’t have a structured way to present the trade-off between economic efficiency and social benefit.

4.5.2 Conflict Detection: After

The system flags the conflict automatically: “BRIA shows marginal economic return (1.2:1). EQIA shows significant equality benefits (fuel poverty reduction, disability access). These findings require ministerial trade-off decision.”

Sarah presents a structured options paper:

- Option A: Proceed (prioritise social benefit over economic efficiency)
- Option B: Modify scope (reduce costs, reduce benefits proportionally)
- Option C: Reject (economic case insufficient)

The minister makes an informed political judgement with full visibility of trade-offs.

4.6 Multi-Agent Architecture

The system uses a supervisor-worker architecture powered by LangGraph. Each agent has a specific role and bounded authority.

Agent Count — To Be Determined Through Development

The exact number of agents will emerge through development and user testing. There could be **dozens or even hundreds of specialised agents**—we won’t know until we build. I designed the architecture for flexibility: we can add, modify, or decompose agents as requirements become clear.

Scaling Pattern: Start with 10–15 core agents in MVP, then specialise and decompose based on user feedback. Complex policies may spawn additional specialist agents dynamically.

Example Initial Agent Categories (illustrative, not exhaustive):

Research Agents (multiple agents per category)

- Qualitative Research agents — Stakeholder submissions, consultation responses, case studies
- Quantitative Research agents — Government statistics, datasets, numerical evidence
- Academic Research agents — Peer-reviewed papers, systematic reviews
- Policy Research agents — Government publications, international comparisons

Analysis Agents

- Conflict Detection agents — Cross-assessment contradictions, trade-offs
- Gap Analysis agents — Missing evidence, data limitations
- Bias Detection agents — Geographic, linguistic, temporal biases
- Legal Trigger agents — Statutory requirements, compliance checking

Synthesis Agents

- Summarisation agents — Executive summaries, key findings
- Template Population agents — Cross-assessment evidence transfer
- Report Generation agents — Formatting, plain language, accessibility

Quality Agents

- Citation Verification agents — Source checking, fact verification
- Consistency agents — Cross-document coherence, terminology
- Supervisor agents — Orchestration, task allocation, quality gates

Agent designs will be validated through user testing with policy professionals during the accelerator phase, ensuring the architecture matches real-world Scottish Government workflows.

This architecture is not theoretical—it builds on production-proven components. But how does it compare to what already exists in the market?

Chapter 5

Competitive Landscape

Sarah's problems are not unique—and neither are the tools that claim to solve them. This chapter examines what exists, what's missing, and why no current solution addresses the integrated impact assessment challenge.

5.1 Ten Tools Analysed

I conducted comprehensive competitive analysis of AI research and impact assessment tools across four categories.

The **AI systematic review space** is well-developed. Elicit offers AI-powered literature screening with up to 80% time savings. Consensus, an AI search engine indexing 220 million scientific papers, includes a Consensus Meter showing scientific agreement levels. Scite, a citation analysis platform serving 2 million users, provides access to 1.4 billion citations classified as supporting or contrasting. Semantic Scholar, from the Allen Institute for AI, provides free access to over 200 million papers. These tools excel at academic research but aren't designed for government policy workflows.

Multi-agent research systems demonstrate the architecture I'm building. LangChain's OpenDeepResearch is an open-source supervisor-worker pattern that orchestrates multiple specialised agents. Perplexity, an AI-powered research assistant with 22 million+ monthly active users and an \$18 billion valuation in 2025, shows market appetite for AI-powered research through iterative search refinement.

For **trade-off analysis**, Agent4Debate, an academic prototype demonstrating AI-powered structured argumentation, achieved human-level debate performance—validating the concept of AI agents presenting opposing viewpoints to surface hidden assumptions.

The **impact assessment software** market is surprisingly thin. OneTrust and TrustArc, enterprise data protection and DPIA platforms, focus on corporate compliance rather

than integrated government impact assessment. No comprehensive platform exists for the full spectrum of government impact assessments.

5.2 Why Existing Tools Don't Solve This

All existing tools are single-type specialists: OneTrust does only DPIAs; Elicit does academic research, not policy workflows; TrustArc focuses on corporate compliance. No tool integrates multiple assessment types across the full government IA landscape.

No evidence reuse: The challenge states “Templates not pre-populated or transferred from one IA to another.” No existing tool solves this.

Different markets: Elicit serves academic researchers; OneTrust serves corporate compliance. Government procurement requires different relationships, certifications, and pricing models. My focus on integrated impact assessment for UK government is an advantage: I'm building exactly what's needed rather than adapting a general-purpose tool.

5.3 Unique Position

Based on comprehensive competitive analysis, this is the first integrated multi-type IA platform powered by agentic AI.

Capability	This Project	Best Competitor
Multi-type IA integration	✓ 8 types	× Single type only
Evidence reuse across assessments	✓	× None offer this
Multi-agent research orchestration	✓	✓ Elicit (single workflow)
AI debate for trade-offs	✓	× Agent4Debate (research only)
Government workflow integration	✓	~ OneTrust (enterprise)
Multilingual research	✓ Neural MT	× English focus

Table 5.1: Competitive Capability Comparison

5.4 Blue Ocean Strategy

I'm not competing in an existing market. I'm creating a new category: integrated government impact assessment with no direct competitors, addressing the unmet need articulated

by the CivTech challenge. 10M+ users across AI research platforms prove demand for evidence-based decision support; zero integrated multi-type IA platforms prove I address a genuine market gap.

The opportunity is clear. What have I already built to seize it?

Chapter 6

Design Foundations

The competitive analysis demonstrates no existing solution addresses the integrated impact assessment challenge. This clarity about market positioning enabled focused design work during the exploration phase. This chapter shows that design work—the foundations that enable development to begin immediately.

6.1 Design Status

The exploration phase produced comprehensive design foundations that enable immediate building when the accelerator begins.

The wireframe specifications total specifications totaling over 69,000 words across all nine workflows. These are not rough sketches but complete screen-by-screen specifications with component architecture, state management patterns, and API endpoint definitions. Accessibility requirements (WCAG 2.1 AA) are built in throughout. When development begins, there are no design questions to answer—only code to write.

Technical validation through build experiments has confirmed the architecture works. The shaden/ui component library meets government accessibility standards, Radix UI primitives pass screen reader testing, and the Next.js application architecture is established with routing and state management. Mock data based on Sarah’s housing retrofit scenario is prepared for demonstration. These are not theoretical choices; they are working code that proves the approach.

Everything is ready for a developer to start building on day one.

This foundation means the accelerator starts building and testing, not designing. Weeks 1–2 focus on user research to validate priorities, then development proceeds immediately against complete specifications. By week 4, the screening prototype is testable. Week 6 delivers evidence gathering with RAG integration. Week 8 adds impact analysis with debate agents. Week 10 achieves end-to-end BRIA workflow. Week 12 completes report

generation. Testing runs throughout—not just at the end—with each milestone validated before the next begins. Weeks 13–15 deliver iteration and refinement based on cumulative feedback.

6.2 Planned Functionality

The accelerator will deliver working implementations of four core workflows that together enable end-to-end impact assessment.

Screening is where every assessment begins. Sarah describes her policy and the system automatically identifies which of the eight IA types are legally required, each with a confidence score and explanation. She can override any recommendation—the system logs her rationale for the audit trail, but the decision remains hers. This eliminates the most common failure mode: missed legal triggers discovered late in the process.

Policy teams explicitly requested this capability. Recent Scottish Government research found that “multiple teams separately suggested that having an interactive tool for IAs would help... an interactive online tool where you could input information about your policy and this would guide you in assessing which IAs you would be required to complete” (Policy Teams’ Experiences Report, August 2025). The screening workflow delivers exactly what policy professionals asked for, validated through 14 interviews with actual IA practitioners.

Evidence Gathering transforms the most time-consuming phase. Sarah describes what evidence she needs in natural language, and the system searches academic databases, government repositories, and international sources simultaneously. Results come back ranked by relevance with explanations—“This source addresses retrofit costs in similar climate conditions”—so she can evaluate the AI’s reasoning. Quality indicators show publication date and source credibility. One click imports a source to her personal evidence library. The system finds more; she curates what matters.

Impact Analysis surfaces the trade-offs that assessments exist to illuminate. The system visualises conflicts between objectives—economic efficiency versus social benefit, business burden versus environmental protection—and flags where different IA types reach different conclusions about the same policy. It generates options for ministerial decision-making, each with clear implications, and breaks down impacts by protected characteristic so nothing is hidden.

Report Generation produces the final output from the single evidence base that has served all assessment types throughout. The system applies consistent formatting across government templates, generates plain language summaries for non-specialist readers, and ensures full accessibility with screen reader compatibility and plain English

alternatives. One policy description, one evidence gathering effort, one analysis process—eight compliant assessment documents.

Each workflow will be tested with policy professionals as it reaches completion—screening in Week 4, evidence gathering in Week 6, and so on—with success measured against the validation criteria below.

6.3 Wireframes and Design Specifications

Wireframe specifications cover all nine workflows totaling approximately 70,000 words. Key screens include Screening Dashboard, Evidence Explorer, Conflict Detection, and Report Preview. Specifications define component architecture, accessibility requirements (WCAG 2.1 AA), state management patterns, and API endpoints.

Full wireframe specifications available on request.

6.4 User Testing Plan

Testing with real policy professionals determines whether the system delivers value. Five to seven participants from different roles—senior policy advisors, junior researchers, specialist leads, managers with approval authority—work through realistic tasks using the prototype. Task-based usability testing reveals where users struggle; System Usability Scale scoring provides quantitative comparison against industry benchmarks; qualitative feedback explores the questions that matter most: whether integration actually reduces fragmentation, whether users trust AI suggestions enough to act on them, and whether trade-offs are presented clearly enough for confident decision-making.

The approach is iterative from the start. Weekly updates based on user feedback mean problems are fixed before they compound. Features are enhanced progressively as user confidence builds—starting simple, adding complexity only where testing shows it helps rather than hinders.

6.5 Validation Protocol

Clear success and failure criteria ensure honest assessment of whether the system delivers value. These thresholds will be applied during user testing from Week 4 onward and refined throughout the accelerator based on Weeks 1–2 user research.

6.5.1 Success Criteria

The system succeeds if policy professionals validate improvement across three dimensions:

Quality Improvement: At least 80% of users rate evidence quality as “better” or “much better” than current process. This is the primary metric—the entire premise is that AI assistance produces higher-quality assessments.

Usability: System Usability Scale (SUS) score of 68 or above, which represents “above average” usability. Users must be able to complete core workflows without extensive training or support.

Task Completion: At least 70% of users successfully complete core tasks (screening, evidence search, conflict review) without assistance. The system should reduce friction, not add complexity.

6.5.2 Failure Criteria (Pivot Triggers)

These thresholds indicate fundamental problems requiring significant course correction:

Quality Failure: Fewer than 50% of users rate evidence quality as improved. If the majority don’t see quality benefits, the core value proposition is invalid.

Usability Failure: More than 3 critical usability blockers per workflow, or SUS score below 50 (“poor”). If users can’t navigate the system, no amount of AI capability matters.

Trust Failure: Users unable to complete core tasks within 2× expected time, or explicit rejection of AI suggestions without review. If policy professionals don’t trust the system, they won’t use it.

6.5.3 Methodology

Testing uses established UX research methods: think-aloud protocol during task completion, post-task questionnaires (including SUS), and semi-structured interviews exploring trust, usefulness, and workflow fit. Sessions are recorded for detailed analysis. Results are tracked quantitatively and reported transparently—including failures.

The design foundations are ready. The testing framework is defined. Now: how does this translate into a 15-week build plan?

Chapter 7

How I’ll Build It

This chapter addresses the core elements of project planning:

Element	Coverage
Objectives	Chapter 1: Why This Matters
Scope	Section 7.3: Scope Boundaries
Timeline	Section 7.5: NOW/NEXT/LATER Roadmap
Resources	Chapter 11: Team Approach
Risks	Chapter 8: Addressing Concerns
Communication	Section 7.4: Stakeholder Communication
Human-AI Synergy	Section 7.2: Human-AI Synergy

Table 7.1: Project Plan Elements

7.1 Agile Approach

Impact assessment sits in the **complex domain** of the Cynefin framework. This distinction matters: complicated problems (assembling a car engine) yield to expert analysis and good practice—you can specify requirements upfront and execute a plan. Complex problems have emergent properties where solutions are only obvious in hindsight.

Why is impact assessment complex? User needs evolve through system interaction. Evidence requirements emerge as research reveals gaps. Stakeholder reactions to findings are unpredictable. Cross-assessment dependencies create feedback loops that cannot be anticipated. A traditional waterfall approach—analyse everything upfront, design the solution, build it, test at the end—would fail because the requirements cannot be fully specified before development begins.

In the complex domain, the appropriate pattern is **probe-sense-respond**: build small increments (probe), gather user feedback (sense), then adapt (respond). This produces

emergent practice rather than predetermined solutions. I use short feedback loops (weekly user contact, fortnightly demos) and safe-to-fail experiments. The 15-week Gantt chart (Figure 7.2) shows the timeline structure and work streams, but specific sprint content adapts based on what we learn.

7.2 Human-AI Synergy

I designed this system for synergy, not replacement. AI agents and human professionals each contribute what they do best, with clear handoffs and human control at every decision point.

What AI agents excel at:

- **Volume:** Searching thousands of sources across languages and jurisdictions simultaneously
- **Synthesis:** Finding patterns and connections across documents that humans would miss
- **Consistency:** Applying the same quality standards to every piece of evidence, every time
- **Speed:** Processing in minutes what would take humans days or weeks

What humans excel at:

- **Judgement:** Weighing trade-offs between competing policy objectives
- **Context:** Understanding political realities, stakeholder sensitivities, and timing
- **Accountability:** Standing behind decisions and explaining them to ministers and the public
- **Ethics:** Recognising when evidence points to uncomfortable conclusions that still must be reported

The architecture ensures humans remain in control. AI agents surface evidence, identify conflicts, and draft analysis—but policy professionals approve every output, can override any recommendation, and see full audit trails of how conclusions were reached. This is human judgement amplified, not automated away.

7.3 Scope Boundaries

Clear scope prevents creep and maintains focus.

In Scope (Accelerator):

- Four core workflows: Planning, Evidence Gathering, Impact Analysis, Report Generation
- Multi-agent backend with LangGraph orchestration and RAG evidence retrieval
- Integration of 1–2 IA types with BRIA as primary (EQIA as secondary if Week 10 progress supports)
- Pilot with 3–5 Scottish Government departments
- Working prototype tested with real policy professionals
- Validation protocol with explicit success/failure criteria

Out of Scope (Accelerator):

- Production deployment or full security accreditation
- Integration with existing Scottish Government IT systems
- International expansion or non-UK regulatory frameworks
- Voice interface or accessibility features beyond WCAG AA

Scope changes require explicit approval and impact assessment on timeline and resources.

7.4 Stakeholder Communication

Fortnightly sprint demos bring the CivTech team and policy professional testers together for live demonstrations of working software. Each session follows a prepared agenda focusing on the most valuable increments, with feedback collected immediately for backlog refinement and recordings available for asynchronous viewing.

Between demos, weekly progress tracking in a shared project board keeps the CivTech team informed against sprint goals. Monthly written reports summarise achievements, user feedback themes, risks, and priorities for the next period. Blockers or pivot triggers requiring stakeholder input are escalated immediately.

Throughout, communication emphasises outcomes achieved and user feedback received—not just activities completed.

7.5 NOW / NEXT / LATER Roadmap

A roadmap shows *what we aim to achieve*—our ambition and intended outcomes. Sprint plans detail *how we get there*—specific activities and tasks. This separation maintains strategic focus while allowing tactical adaptation. Following GOV.UK Service Manual guidance, the roadmap is revisited after each major milestone based on what we learn.

7.5.1 NOW (Accelerator Weeks 1–6)

Learning Objective: Validate core workflow integration and evidence gathering assistance

Deliverables:

- Complete user research with 5–7 policy professionals
- Screening workflow with legal trigger identification
- Evidence gathering with multi-source search (RAG system operational)
- Initial multi-agent orchestration with LangGraph (research and synthesis agents)
- Basic evidence library and cross-assessment viewing
- User testing and iteration

Exit Criteria: Users can screen a policy and gather initial evidence faster than current process, with equal or better quality

Week 6 Decision Point: Review progress against timeline. If behind schedule, reduce scope to single workflow depth over breadth. If on track, proceed with full NEXT phase plan.

7.5.2 NEXT (Accelerator Weeks 7–12)

Learning Objective: Validate conflict detection and report generation

Deliverables:

- Impact analysis with trade-off visualisation
- Debate agents for conflict detection and resolution suggestions
- Cross-assessment conflict flagging
- Report generation for BRIA (primary IA type)
- Full multi-agent pipeline operational end-to-end
- User testing and iteration

Exit Criteria: Users can complete BRIA end-to-end with measurable quality and time improvements

Week 10 Decision Point: Critical assessment of multi-agent backend stability and user feedback. Determines LATER phase scope:

- If behind: Focus exclusively on primary goal (BRIA polished and production-ready)
- If on track: Proceed with secondary goal (add EQIA integration)
- If ahead: Target stretch goal (add 1–2 additional types)

7.5.3 LATER (Accelerator Weeks 13–15 + Post-Accelerator)

Learning Objective: Validate multi-type integration and institutional learning

Deliverables follow a tiered approach based on Week 10 progress review:

Primary Goal (Must achieve):

- BRIA end-to-end with full multi-agent backend operational
- Evidence reuse library with cross-assessment knowledge transfer
- Portfolio dashboard demonstrating single-type integration
- System validated by policy professionals with measurable quality improvements

Secondary Goal (If on track at Week 10):

- Add EQIA (Equality Impact Assessment) integration
- EQIA shares significant evidence overlap with BRIA (protected characteristics, consultation, impact analysis)
- Demonstrates multi-type integration capability with 2 assessment types
- Cross-type conflict detection operational between BRIA and EQIA

Stretch Goal (If significantly ahead of schedule):

- Add 1–2 additional IA types (CRWIA, FSDA, or DPIA)
- Full 8-type integration remains a 6–12 month post-accelerator goal
- Accelerator focuses on proving the concept with 1–2 types delivered excellently

Rationale: Better to deliver 1–2 assessment types with production quality, comprehensive testing, and validated user value than 8 types partially complete. The primary accelerator objective is validating the integrated workflow architecture and multi-agent orchestration with real policy professionals—not maximizing type coverage. Once the model is proven with BRIA (and ideally EQIA), expanding to additional types becomes a scaling exercise rather than a validation challenge.

Exit Criteria: Primary goal complete with quality metrics validated (80%+ users rate evidence as better, SUS score 68+). Secondary goal demonstrates multi-type integration works. Stretch goal achieved only if it doesn’t compromise primary/secondary quality. Clear documentation of remaining work for post-accelerator development.

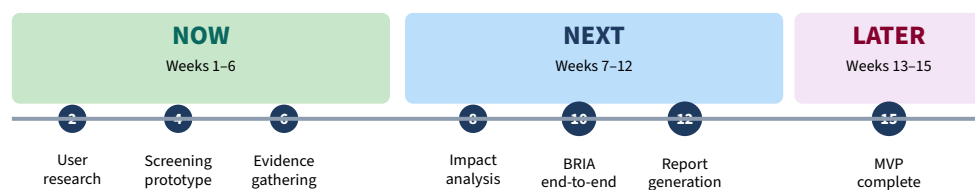


Figure 7.1: 15-week accelerator roadmap with key milestones

7.6 Key Milestones

7.7 Momentum Already Built

I’m not starting from zero. The exploration phase produced:

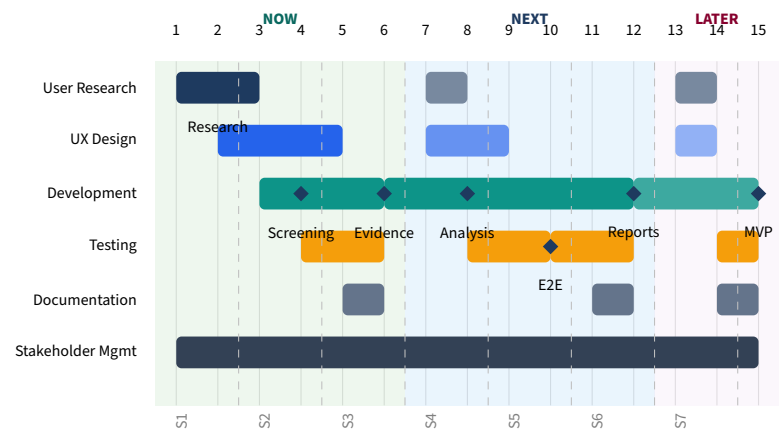


Figure 7.2: 15-week Gantt chart showing work streams and sprint boundaries (S1–S7)

Week	Milestone	Validation Method
2	User research complete	5+ interviews documented
4	Screening prototype	User testing (3+ users)
6	Evidence gathering + RAG	Time comparison vs manual
8	Impact analysis + debate agents	Trade-off clarity scoring
10	BRIA end-to-end	Quality comparison vs manual
12	Report generation complete	User satisfaction (SUS 70+)
15	MVP complete (1–8 IA types)	Ready for pilot deployment

Table 7.2: Key Milestones and Validation Methods

- **Research:** 198 pages of persona documentation (15 professionals across four tiers), 147 user stories across eight workflows, 12 journey maps, 10 competitive tools analysed
- **Design:** Eight workflow specifications (~4,000 lines), shadcn/ui component library selected
- **Prototype:** React/Next.js foundation established, core components implemented, demo data prepared

The accelerator builds on this foundation rather than starting with a blank screen.

Ambitious timelines and comprehensive design foundations create confidence—but what could go wrong, and how will I handle it?

Chapter 8

Risks and Compliance

When Sarah submits her impact assessment to ministers, she needs absolute confidence that every citation is real, every statistic is accurate, and every trade-off has been properly analysed. The consequences of failure are severe: judicial review if legal requirements are missed, ICO enforcement with fines up to £17.5M if data protection is breached, and erosion of public trust if policy decisions are based on flawed evidence.

I've designed this system with these stakes in mind. Every technical choice, every architectural decision, reflects careful risk analysis. Here's how I ensure the system is trustworthy enough for government use.

8.1 Ensuring AI Accuracy

The greatest technical risk in any AI system is hallucination—generating plausible but incorrect information. In a government context, a hallucinated citation could undermine an entire impact assessment. This risk is why shadow AI usage is so dangerous: ChatGPT has no accountability, no source verification, no audit trail.

Primary Defence: Human Review at Every Phase

My approach addresses this risk through a fundamental architectural principle: **no evidence enters any assessment without explicit human approval**. AI agents find, rank, and summarise potential evidence—but humans review and approve every source before it becomes part of the assessment. This isn't a final check; it's built into the core workflow.

The 7-phase system model enforces human checkpoints throughout, creating a chain of approval that AI cannot bypass.

In **Screening**, AI identifies potential legal triggers across all eight IA types, but the human reviews and confirms which assessments are actually required, with rationale

logged for the audit trail. In **Planning**, AI suggests evidence gaps and research priorities, but the human approves the research plan before any evidence gathering begins. This prevents wasted effort on irrelevant searches.

Evidence Gathering is the critical phase for hallucination control. AI retrieves candidate sources from multiple databases, but *the human reviews every single piece of retrieved evidence*—approving, rejecting, or flagging each source. Only human-approved evidence proceeds to analysis. This is not a spot check; it is comprehensive review of everything the AI finds.

In **Impact Analysis**, AI synthesises the approved evidence and identifies conflicts between assessment types, but the human reviews the analysis, validates the trade-offs, and approves conclusions before they become part of the assessment. In **Report Generation**, AI drafts report sections from approved content, but the human reviews every section before finalisation—no AI-generated text goes to ministers without human approval.

Quality Assurance brings specialist review: Aisha reviews EQIA conclusions, Raj reviews BRIA cost estimates. The system flags items for attention, but humans make all decisions. Finally, in **Approval**, the designated approver provides final sign-off with full audit trail showing every human decision throughout the process.

Hallucination risk is therefore contained at the evidence gathering phase. Even if AI retrieves an incorrect or fabricated source, it cannot enter the assessment without human approval. The system produces candidates; humans curate what gets used.

Supporting Technical Layers

Human review is the primary defence. Technical measures make that review more efficient and effective:

GraphRAG for better retrieval: Microsoft Research demonstrated 3–4× accuracy improvement over traditional methods for complex queries. Rather than generating answers from training data, the system retrieves evidence from verified sources. Humans review better candidates, not more candidates.

Full citation transparency: Every AI-retrieved source includes URL, publication date, and author. Sarah can click any citation and see the original. If the system cannot find supporting evidence, it says so explicitly rather than inventing a reference—enabling rapid human verification.

Confidence scoring for prioritisation: Sources rated below 70% relevance are flagged with warnings: “This source may not fully support the claim. Please verify carefully.” Humans can then prioritise their review effort on uncertain items while still reviewing everything.

Result: Defence in Depth

The combination of mandatory human approval at every phase plus technical measures that improve retrieval quality and enable efficient verification creates defence in depth. Modern RAG-based models achieve sub-1% hallucination rates (Vectara Hallucination Leaderboard, November 2025)—but even that small risk is caught by the human review gate before it can affect the assessment. Compare this to shadow AI tools where hallucinated content goes directly into ministerial briefings with no verification whatsoever.

8.2 Protecting Sensitive Data

Impact assessments often contain sensitive information: unpublished policy positions, personal data from consultations, commercially confidential business cost estimates. A data breach would be catastrophic—not just the £17.5M maximum GDPR fine, but the destruction of trust that would set back government AI adoption by years.

Enterprise Model Deployment

A critical distinction: this system uses enterprise LLM deployments, not consumer AI tools. When policy professionals use ChatGPT directly, their prompts go to OpenAI's servers, may be retained, and historically were used for model training. Enterprise Azure OpenAI is fundamentally different.

Microsoft explicitly commits that customer data is not used for model training unless explicitly opted in. Scottish Government data will never train OpenAI's models—full stop. The system runs on Provisioned Throughput Units (PTU), which provide dedicated, isolated compute capacity reserved specifically for Scottish Government workloads. This is not shared infrastructure where government prompts mingle with commercial queries; it is reserved capacity with both performance guarantees and additional data isolation.

Enterprise customers can also opt out of the default 30-day prompt retention that exists for abuse detection. Through the Modified Abuse Monitoring approval process—routinely granted to government organisations—prompts are processed and immediately discarded. Zero data retention beyond request completion. And critically, data is never shared with OpenAI or other third parties. The models run entirely within Microsoft's Azure infrastructure under Microsoft's data protection commitments, not OpenAI's consumer terms.

This is the opposite of shadow AI risk. When Sarah uses ChatGPT in her browser, she has no control over data handling—her prompts may be stored, analysed, and used for training. When she uses this system, every interaction is governed by enterprise

agreements, UK data residency requirements, and explicit no-training commitments. The difference is not just technical; it is contractual and legally enforceable.

Infrastructure Security

The system runs entirely on Azure UK regions (UK South and UK West), which are G-Cloud (UK Government’s cloud computing framework) certified and compliant with OFFICIAL-SENSITIVE (UK government’s restricted information classification level) requirements. Data never leaves UK jurisdiction. All information is classified on ingest, with access controls enforced by classification level. Role-based access follows the principle of least privilege: James can see evidence relevant to his research tasks, but not the ministerial briefing that Emma is preparing.

Every access is logged in an immutable audit trail with real-time anomaly detection. If someone attempts to access data outside their normal pattern, the system flags it immediately. And through Microsoft’s Customer Lockbox, no Microsoft engineer can access Scottish Government data without explicit approval from a designated government administrator.

8.3 Building User Trust

The most likely reason for project failure isn’t technical—it’s human. Policy professionals may resist AI-assisted working, viewing it as threatening their expertise or adding complexity to already-stressful deadlines. If Sarah and her colleagues don’t trust the system, they won’t use it, and the investment will be wasted.

This is why I start with the manual-first approach. Before any AI agents are introduced, users experience the integrated workflow: eight assessment types working together, evidence shared across templates, legal triggers identified automatically. They see immediate value without any AI complexity. Only then do I add agent assistance, beginning with evidence gathering where the benefits are clearest and the risks lowest.

Co-design from Week 1 means policy professionals shape the system rather than receiving a finished product. When Sarah identifies a workflow that doesn’t match how her team actually works, we adjust. When James finds the interface confusing, we simplify. This creates ownership: it’s their system, not something imposed on them.

I’ll also establish a champions network—early adopters who support their colleagues through the transition. These aren’t IT specialists; they’re respected policy professionals who can demonstrate value to their peers. Combined with comprehensive, role-specific training and transparent sharing of quality metrics, this approach transforms potential resistance into genuine adoption.

8.4 Maintaining Human Judgement

There's a subtler risk than outright resistance: over-reliance. If policy professionals start trusting AI suggestions without applying their own judgement, assessments might be technically complete but analytically weak. The system would produce documents, but not the insights that inform good policy.

I guard against this through consistent messaging: AI assists, humans judge. The interface reinforces this at every step. When the system suggests evidence, it explains why: "This source is relevant because it addresses retrofit costs in similar climate conditions." Sarah can evaluate that reasoning. When the system flags a conflict between BRIA and EQIA findings, it presents options but explicitly states: "This trade-off requires ministerial judgement."

Quality metrics track analytical depth, not just completion. Specialists like Aisha continue reviewing assessments, now with more time to focus on complex cases rather than routine queries. And ongoing training helps users understand when to trust AI recommendations and when to apply human judgement—a skill that becomes more valuable as AI tools proliferate across government.

8.4.1 Mechanical Safeguards Against Over-Reliance

Beyond messaging, the interface includes structural friction that prevents mindless acceptance:

Mandatory Edit Fields: Every AI-generated summary includes a "Your Assessment" field that must be completed before proceeding. Sarah cannot simply click "Accept All"—she must articulate her own view, even if it's "I agree because..."

Forced Reflection Points: At key decisions (evidence inclusion, conflict resolution, risk rating), the system presents a brief checklist: "Have you considered alternative interpretations? Have you consulted the specialist guidance? Are you confident in this judgement?" These take 10 seconds but interrupt autopilot mode.

Confidence Calibration: The system periodically asks Sarah to rate her confidence before revealing AI analysis, then shows comparison. If she's consistently under- or over-confident relative to specialist review, she receives tailored guidance. This builds metacognition about when to trust her own judgement versus when to seek input.

Variation in AI Presentation: Not every suggestion looks the same. Some are presented as strong recommendations, others as weak possibilities, others as "Here are three options—you decide." This prevents Sarah from developing a single mental shortcut for all AI content.

Audit of Human Reasoning: The immutable log captures not just what Sarah decided but her stated rationale. Quality reviews can identify patterns of superficial reasoning (“Agreed with AI suggestion”) versus genuine analysis (“AI suggested X but I chose Y because of local context Z”). This creates accountability without micromanagement.

These safeguards are not punitive. They are designed to support the development of effective human-AI collaboration skills—skills that will become increasingly valuable as AI tools proliferate across government.

8.5 Regulatory Compliance

The system is designed to comply with the UK Government AI Playbook (February 2025), which establishes ten principles for responsible AI use in government. Full compliance validation will occur during the accelerator with Scottish Government security teams. This isn’t a compliance checkbox; it’s the foundation of trustworthy AI.

Principle	Implementation
1. Know AI and limitations	Three Hows Framework explains capabilities and risks
2. Use AI lawfully, ethically	GDPR compliance, EU AI Act alignment
3. Use AI securely	Azure UK regions (G-Cloud certified), least-privilege agents
4. Meaningful human control	Human approval required at all 7 phases; no evidence used without human review
5. Manage AI lifecycle	Full lifecycle planning including decommissioning
6. Right tool for the job	“When NOT to use AI” guidance
7. Quality data	Evidence quality scoring, gap identification
8. Iterate and improve	Continuous learning, quarterly enhancements
9. Prepare and upskill	Comprehensive training, 3-tier support
10. Address wider harms	EQIA for system itself, environmental monitoring

Table 8.1: UK AI Playbook Compliance

The system also directly addresses the seven risks identified in Scottish Government IA guidance. Where current practice treats assessments as tick-box exercises done at the last minute, this system makes them valuable inputs to policy design from the earliest stages. Where current consultations suffer from poorly expressed questions and low response rates, AI-assisted design improves both question quality and stakeholder targeting. Where current assessments reach inconclusive conclusions because trade-offs are hidden, this system surfaces conflicts explicitly and structures options for ministerial decision.

The goal is not just compliance, but transformation: from disconnected bureaucratic burden to integrated analytical asset.

Chapter 9

Commercial Viability

The system works. The risks are managed. Now: is this a sustainable business? This chapter presents the commercial case—starting with Scottish Government and growing from there.

9.1 The Growth Story

This project begins with Scottish Government and grows from there. The strategy is deliberately conservative: prove value with one customer before expanding, build a reference case before selling to others, demonstrate results before making promises.

Scottish Government produces several hundred impact assessments per year across eight types. If I can help Sarah and her colleagues produce better assessments faster, that success creates the foundation for everything that follows. A working system in Scottish Government is worth more than a hundred slide decks—it's proof that the technology works, the approach is sound, and the team can deliver.

9.2 Market Expansion

Once Scottish Government is successful, the expansion path is clear. UK Government operates 24 ministerial departments plus over 100 executive agencies, all producing impact assessments under similar requirements. The devolved administrations in Wales and Northern Ireland have comparable needs. Local authorities—418 across the UK—face the same challenges at smaller scale. Each successful deployment adds evidence to the reuse library, making the platform more valuable for everyone.

The global market for AI in government and public services is substantial: \$26.4 billion in 2025, growing to \$135.7 billion by 2035 according to Future Market Insights. EU member states have similar impact assessment requirements under EU law. Commonwealth

nations share legal traditions that make the platform readily adaptable. The opportunity is genuinely international, but I won't pursue it until the UK foundation is solid.

Competitive positioning is strong because no integrated multi-type impact assessment platform exists globally. Elicit does academic research; OneTrust does privacy compliance; neither integrates eight assessment types or enables evidence reuse across them. First-mover advantage in a critical government function, combined with platform economics where each customer adds value for all others, creates defensible market position.

9.2.1 Procurement Reality and Go-to-Market Strategy

I am not relying on optimistic assumptions about government sales cycles. New government technology vendors typically face 18-month minimum sales cycles for direct tenders. My approach acknowledges this reality through the CivTech pathway.

CivTech Accelerator Funding and Deployment

The CivTech accelerator provides a commercial contract valued at £350,000 + VAT that funds both the 15-week development programme and initial deployment with Scottish Government. This structured pathway eliminates the typical Year 1 challenge for government technology startups: how to fund development before securing first revenue.

Successful accelerator completion leads directly to deployment across 3–5 Scottish Government departments, establishing the reference case that enables UK market entry. This is not speculative revenue dependent on winning competitive tenders—it is the designed outcome of the CivTech programme structure.

9.3 Scottish Economic Value

This project creates high-value employment in Scotland: solo founder during accelerator, growing to two roles in year one (senior AI engineer, full-stack developer), three to four in year two (adding policy/UX specialist and sales lead), and five to eight Scotland-based roles by year three. These positions combine AI and ML engineering with policy domain expertise—exactly the kind of high-skill jobs that Scotland's economic strategy prioritises.

Beyond direct employment, the project develops AI skills across Scottish Government and demonstrates that Scotland can be an early adopter of trustworthy AI. When government successfully deploys an AI system for a critical function, it signals to the broader ecosystem that Scotland is a place where AI talent can do meaningful work. This aligns directly with Scotland's AI Strategy, which calls for “trustworthy, ethical and inclusive AI”—precisely what the human-in-the-loop approach delivers.

9.4 Procurement and Sustainability

I designed the platform for government procurement pathways from day one: data ownership stays with government, export formats are standard, and open-source components reduce dependency. I built in exit planning because governments rightly worry about vendor lock-in. Data ownership stays with government. Export formats are standard. Open-source components reduce dependency. If Scottish Government ever wants to move to a different provider, they can take their data and their evidence library with them.

This isn't just good ethics; it's good business. Governments choose vendors they can trust for the long term. Demonstrating commitment to their interests, not just my own, builds the kind of relationship that leads to contract renewals and referrals.

9.5 Platform Vision

Impact assessments are the beginning, not the end. The deeper opportunity is transforming how large organisations make evidence-based decisions.

9.5.1 The Bigger Opportunity

Today, governments and enterprises make critical decisions using static documents—Word files and PowerPoint decks that are immediately out of date, impossible to personalise, and disconnected from source evidence. A minister receives a 50-page briefing but cannot ask follow-up questions. An executive gets a due diligence report but cannot explore alternative scenarios. The document sits inert on a desk while the world it describes keeps changing.

The platform I am building enables something fundamentally different: living documents that are dynamic, interactive, evidence-connected, and personalised—yet embedded in systematic organisational workflows. Users can chat with their assessments, asking questions like “What does the Swedish research say about this?” or “Show me the trade-off between cost and equality impact.” They can explore data through personalised dashboards that present information the way each person thinks, not a one-size-fits-all format. They can drill into source material instantly, seeing exactly which evidence supports each conclusion.

This is not speculative technology. It is the architecture I am building for impact assessments, applied more broadly. The same multi-agent orchestration that gathers evidence across databases and languages can gather evidence for any complex organisational

question. The same conflict detection that identifies tensions between business and equality impacts can identify tensions between any competing objectives. The same report generation that produces accessible, consistent documents can produce any document type an organisation needs.

9.5.2 Modular Architecture for Expansion

The system is designed for extension from the ground up. The supervisor-worker architecture that orchestrates research and analysis agents can incorporate new specialist agents without rebuilding the core—each new capability plugs into the existing orchestration layer rather than requiring parallel infrastructure.

The first natural extension is AI voice agents for stakeholder consultations. Rather than transcribing hours of interviews manually, voice agents capture structured input directly, categorise responses, and identify themes. The technology exists; I am already developing voice agents with Microsoft's Technology Centre London for other applications, and the same capability transfers directly to policy consultation.

From there, the same evidence synthesis that produces impact assessments can produce document intelligence agents for ministerial briefings and parliamentary questions—other document types that policy professionals create repeatedly with similar evidence-gathering and analysis requirements. Scenario modelling agents extend the analysis capability further, enabling interactive exploration of how changing assumptions affects outcomes rather than static options papers. And real-time monitoring agents connect assessments to actual policy outcomes, so each new assessment learns from what happened after the last one was implemented.

Each module adds capability to the platform without requiring customers to adopt everything at once. A department can start with impact assessments, prove value, and expand to adjacent use cases as confidence builds. The architecture supports this incremental adoption because each agent type shares the same orchestration infrastructure, evidence management, and quality controls.

9.5.3 Phased Market Expansion

The path from impact assessment tool to enterprise platform follows natural validation stages, each building on the last:

Years 1–2: Government Impact Assessments

The focus is proving the model works. Scottish Government is the ideal first customer: clear pain articulated in the CivTech challenge, legal requirements that mandate completion,

direct access to policy professionals through the accelerator, and the credibility that comes from meeting government standards for accuracy and accountability. Success here—measurable quality improvements, user adoption, ministerial satisfaction—creates the foundation for everything that follows.

Years 3–4: Adjacent Government Use Cases

With impact assessments proven, the same architecture applies to adjacent functions that policy teams perform repeatedly: policy development and options analysis, parliamentary questions and FOI responses, ministerial briefings and Cabinet papers, consultation design and response analysis. Each represents the same fundamental challenge—gathering evidence, synthesising analysis, presenting trade-offs—that impact assessments require.

This expansion happens within existing customer relationships. A department already using the platform for impact assessments extends to briefings and parliamentary questions. The evidence library they have built transfers directly; the workflows they have learned apply with modification.

Years 5–7: Enterprise Market Entry

The lessons learned from government—where stakes are highest, scrutiny is intense, and accountability is absolute—transfer to industry. If the platform meets government standards for accuracy and transparency, it exceeds enterprise requirements by default.

Years 8–10: Global Platform

The ambition is to become the standard for complex organisational decision-making worldwide. International government expansion follows the UK model: start with one department, prove value, expand across government. Enterprise adoption deepens as the platform demonstrates value across industries and use cases.

9.5.4 Why This Creates Exceptional Value

Most AI tools are “useful”—they help people with tasks. This platform produces output: completed assessments, finished briefings, published reports. Every output represents hours of professional time and tangible organisational value. The difference matters for how customers think about the investment.

A tool that helps with tasks competes on efficiency: does it save 20% of time or 30%? A platform that produces output competes on capability: can you do things you could not do before? The impact assessment platform enables international evidence at scale, multilingual research, contradiction detection across eight assessment types—capabilities that manual processes cannot match regardless of time invested.

In government, better evidence synthesis means better policy decisions. Fewer missed impacts that lead to judicial review. Fewer contradictions that undermine ministerial confidence. Better outcomes for citizens. The value is measured in policy quality, not just time savings.

In enterprise, the same capability applied to M&A, ESG, or regulatory compliance means faster decisions with better risk coverage. The organisations that make better decisions faster win.

The long-term total addressable market—share of all evidence-based organisational decision-making—extends into trillions. That is not a projection for this business; it is the scale of the problem. Every organisation makes complex decisions based on gathered evidence. Today they do it with static documents, email chains, and manual processes. The opportunity is to transform that fundamentally.

9.5.5 Why Start With Impact Assessments

Government is the hardest market to win. Security requirements are highest. Risk tolerance is lowest. Procurement cycles are longest. If the platform succeeds in government, it succeeds anywhere.

Impact assessments are the ideal entry point within government. The pain is clear and officially articulated in the CivTech challenge. Legal requirements mandate completion—there is no optional adoption decision. Outcomes are measurable—better evidence, consistent standards, successful judicial review defence. The accelerator provides direct access to policy professionals who will shape the product.

Success with impact assessments creates the reference case for everything that follows. A minister who trusts the platform for impact assessments will consider it for briefings. A department that sees quality improvements will expand to other use cases. An enterprise evaluating AI for due diligence will weight government validation heavily.

Impact assessments are the proving ground, not the destination.

The Ambition: Build the platform that transforms how organisations make evidence-based decisions—first in government, then in enterprise, ultimately becoming the standard for complex decision-making worldwide. Not a tool that helps with tasks, but a system that produces the outputs organisations need. Not replacing human judgement, but augmenting it with better evidence, clearer analysis, and transparent trade-offs at a scale humans alone cannot achieve.

Chapter 10

Environmental Strategy

10.1 Direct Environmental Benefits

The project operates with minimal environmental footprint from day one. Fully remote working means zero office space and no commuting emissions—a genuine reduction, not an offset. Cloud infrastructure runs exclusively on Azure UK regions (UK South in London and UK West in Cardiff), which are G-Cloud certified for government use and supported by Microsoft’s renewable energy commitments. AI computation is managed for efficiency: smaller models handle routine operations, prompt caching reduces redundant processing, and batch jobs run during low-carbon grid periods when renewable generation exceeds demand.

10.2 Platform Environmental Impact

The platform’s larger environmental contribution comes through improving the quality of Strategic Environmental Assessments themselves. Better SEAs mean better environmental outcomes from policy. When Sarah gathers evidence for the housing retrofit policy, the system surfaces international research on energy efficiency that UK-focused searches would miss—German Passivhaus standards, Swedish district heating integration, Danish renewable energy mandates. This evidence directly improves the environmental analysis.

Cumulative impact detection identifies where multiple policies affect the same environmental systems. If three different policies all increase demand for construction materials, the system flags this overlap so decision-makers can consider the combined environmental effect rather than treating each policy in isolation.

10.3 Measurable Commitments

Environmental responsibility requires transparency and accountability, not just good intentions.

Renewable Energy Commitment: All computation runs on Azure UK South and UK West data centres. While these facilities physically connect to the UK National Grid (which includes fossil fuels, nuclear, and renewables), Microsoft matches 100% of energy consumption through renewable energy power purchase agreements (PPAs) and Renewable Energy Guarantees of Origin (REGOs) on an annual basis. This contractual matching approach, while not providing 24/7 carbon-free power in real-time, represents Microsoft's commitment towards their goal of 100% carbon-free energy by 2030.

User-Controlled Environmental Dashboard: The platform includes an environmental dashboard that gives users direct control over computational resources. Policy professionals can see the relative environmental impact of different workflow configurations and adjust how deeply AI agents search, how many sources they consult, and which models they use. A quick scoping query uses minimal resources; a comprehensive evidence synthesis uses more. Users make informed trade-offs between thoroughness and environmental impact, rather than having those decisions hidden from them.

The dashboard shows:

- Relative environmental impact per workflow
- Comparison to previous assessments
- Settings to adjust agent depth and breadth
- Model selection options (smaller models for routine tasks)

Year-Over-Year Improvements: I commit to reducing per-assessment emissions through model efficiency gains—smaller models where appropriate, prompt caching to eliminate redundant processing, and batch scheduling during low-carbon grid periods. Quarterly environmental reports with CodeCarbon integration provide detailed monitoring that users and stakeholders can verify.

10.4 Net Zero 2030

The company commits to Net Zero emissions by 2030—an ambitious five-year timeline that reflects the urgency of climate action. This is achievable for a cloud-native, remote-first startup because the footprint is inherently small and the levers for reduction are within direct control.

Scope 1 emissions (direct from owned sources) are near-zero from day one: no offices, no company vehicles, no on-premises data centres.

Scope 2 emissions (indirect from purchased energy) are addressed through Azure UK data centres, where Microsoft matches energy consumption with renewable sources through annual power purchase agreements and renewable energy certificates, working towards their 100% carbon-free energy goal by 2030.

Scope 3 emissions (supply chain) are the challenge. The approach: sustainable procurement, carbon-aware computing that schedules non-urgent work during low-carbon grid periods, and verified offsets from Gold Standard or Verified Carbon Standard projects for residual emissions that cannot be eliminated.

Annual ESG reporting using the Global Reporting Initiative Standards framework provides transparent tracking toward this commitment. The reports will be published publicly even while the company is small, because environmental responsibility requires accountability from the beginning, not promises to act once scale is achieved.

10.5 User-Controlled Environmental Settings

Environmental responsibility should not require users to accept a one-size-fits-all approach. Different assessments have different needs: an urgent scoping query does not require the same depth as a comprehensive analysis for a major policy. The platform gives users direct control over this trade-off.

Research Depth Settings:

Quick Mode searches the top 100 most relevant sources with a single research round. This is appropriate for initial scoping, urgent queries, or simple assessments where extensive international research is unnecessary. Lowest environmental impact.

Standard Mode searches the top 400 sources across multiple databases, with multi-agent research including two rounds of debate to surface contradictions and trade-offs. This is appropriate for most assessments and balances thoroughness with efficiency. Moderate environmental impact.

Comprehensive Mode searches over 1000 sources including international and multilingual research, with three or more debate rounds and deep cross-assessment analysis. This is appropriate for critical policies with significant impacts, major consultations, or assessments likely to face legal challenge. Higher environmental impact due to computational intensity.

Environmental Priority Toggle:

Users can choose between Speed Priority, which uses the latest and fastest models with immediate processing, and Green Priority, which uses energy-efficient models like Gemini Flash and batches processing during low-carbon grid periods. Green Priority may add modest delays to non-urgent work but reduces carbon footprint substantially.

Resource Controls:

Users can set computational resource limits for each assessment. The system adapts research depth automatically to stay within these limits, warning if the limits may compromise quality. This enables departments to manage environmental impact alongside research thoroughness with full visibility.

Before starting each assessment, users see projections of processing requirements and number of sources that will be consulted based on the selected mode (Quick, Standard, or Comprehensive). This transparency enables informed decisions about research depth appropriate to each policy's importance.

Chapter 11

Team, Diversity and Growth

The technology is proven. The risks are managed. The market is clear. Now: who builds this, and how does the team grow?

11.1 Founder Capability

I bring over a decade of AI experience spanning academia, aerospace, entrepreneurship, and enterprise-scale production systems.

Academic: DPhil at Oxford University Big Data Institute, designing end-to-end computer vision pipelines that transformed raw data into structured, explainable outputs. This mirrors what impact assessments require: moving from scattered evidence to actionable analysis.

Industry: At Astroscale, I built production-grade AI for satellite operations in regulator-facing environments where failure was not an option. This taught me to build systems that are traceable, reliable, and trusted.

Entrepreneurial: At Datamise (a separate company I co-founded), I built SettleMise for equal pay claims. I delivered the full stack while working directly with non-technical users (lawyers, union officials). Strong evidence I can do the same for policy professionals.

Current: At SThree, I'm building multi-agent AI systems orchestrating 100+ specialised agents. I'm also developing AI voice agents with Microsoft's Technology Centre London, and RAG systems connecting LLMs to large regulatory datasets. These are exactly the capabilities this project requires: agent orchestration, voice consultations, and evidence-grounded responses.

11.2 Why Solo Founder Works

A solo founder with AI augmentation can move faster than a small team spending time in meetings. The key is quantifying what AI tools actually deliver:

Development Velocity: Studies show AI coding assistants enable developers to complete significantly more tasks on average. The gains are strongest for boilerplate code, test generation, documentation, and debugging. Complex architectural decisions still require human judgement, but targeted use of AI for appropriate tasks accelerates development substantially.

Research Throughput: AI research agents (the same technology I’m building for policy professionals) enable 5–10× faster literature review and competitive analysis. The 198 pages of persona documentation, 147 user stories, and 10-tool competitive analysis in this proposal were produced in three weeks—work that would take a small team months.

Net Effect: Solo founder with AI augmentation delivers substantial output for development and research tasks, without the coordination overhead, meeting time, or communication friction. Decisions are made quickly; user feedback becomes working implementation in days rather than weeks.

This follows Y Combinator’s “billion dollar ten-person company” philosophy: stay lean, avoid coordination overhead, maintain focus. It’s not about doing less with less—it’s about doing more with AI augmentation while keeping decision-making fast and accountable.

This does not mean staying solo forever. Hiring follows product validation and user feedback, not predetermined headcount targets.

Advisory Support Through CivTech: The CivTech accelerator itself provides the advisory infrastructure that a traditional startup would need to build externally. The programme’s mentorship, the challenge sponsor relationship, and direct access to Scottish Government policy professionals give me everything required to refine both product and business case. This is product-focused development at its purest: building with real users from day one, not seeking external advisors who lack domain context. The policy professionals who will use the system are the best advisors for shaping it.

11.3 Hiring Roadmap

Each hire addresses a specific scaling need validated through product development and user feedback. The sequence prioritises building capability—the ability to extend and improve the platform—over advisory roles that can be contracted or accessed through CivTech.

Year 1 (Post-Accelerator):

The first hire, around month six, is a Senior AI Engineer specialising in LangGraph and multi-agent systems. The multi-agent orchestration is the core intellectual property of the platform; as adoption grows, someone needs to scale it from the MVP's 10–15 agents to production systems with dozens or hundreds. This person extends the architecture, improves reliability, and handles the technical complexity that emerges as the platform serves more users and use cases. The hire is technically deep—this is not a generalist role.

The second hire, around month nine, is a Full-Stack Developer with React and Next.js focus and backend capability. Government users need polished, accessible interfaces, and user feedback during the accelerator will generate substantial UI iteration needs. This person builds new workflows, refines existing ones based on feedback, and ensures the frontend matches the quality users expect from government systems.

Year 2:

The third hire is a Policy and UX Specialist—a hybrid role that bridges user needs and product development. The ideal candidate is a former policy professional who has learned UX methodology, or a UX designer with substantial government experience. This person translates what Sarah and her colleagues say they need into product requirements that engineers can build, and tests whether what we built actually solves the problem. User research during the accelerator is contracted to specialist firms; this hire brings that capability in-house as the user base grows.

The fourth hire is a Sales and Commercial Lead. As the platform proves value in Scottish Government and early UK departments express interest, someone needs to manage those relationships, navigate procurement processes, and build the case studies that enable expansion. Government sales requires relationship building and procurement expertise that differs from the founder's technical background.

Year 3:

The team grows to five to eight people, adding additional engineers as platform complexity and user demand increase, customer success capability as the customer base expands, and potentially roles supporting international expansion if demand materialises. Each hire remains product-driven: we add capability when adoption metrics and user feedback demonstrate the need, not when a headcount plan says we should.

UX research during the accelerator comes from contracted specialist firms who conduct user testing, analyse feedback, and recommend improvements. Domain expertise comes directly from policy professional access through CivTech—the users themselves. The

hiring roadmap focuses on building the team that can extend and scale the platform, not on roles that can be accessed through other means during early stages.

11.4 Agile Thinking Demonstrated

This exploration phase demonstrates Agile methodology: user-centred research before code, iterative refinement against the CivTech scorecard, agent validation across six policy scenarios, NOW/NEXT/LATER adaptive planning.

For SettleMise, I worked iteratively with GMB union representatives, building features based on what they actually needed. The first version was basic; each iteration added functionality based on feedback. The result was a system built around their actual workflow.

The accelerator continues this approach: weekly user contact, fortnightly demos, regular retrospectives, safe-to-fail experiments for uncertain features.

11.5 Diversity Commitment

I grew up in a former mining village outside Airdrie, North Lanarkshire. I know first-hand that class background is one of the most overlooked barriers in technology. Socioeconomic diversity is often treated as an afterthought—organisations focus on gender and ethnicity (important) while ignoring the class barriers that prevent talented people from ever reaching the applicant pool. My approach puts socioeconomic inclusion alongside gender, ethnicity, and disability as core priorities.

11.5.1 Inclusive Hiring Practices

The hiring process itself must not create barriers. Blind CV screening removes names, universities, and photographs to reduce unconscious bias—research consistently shows that identical CVs receive different responses based on names that signal gender or ethnicity. Structured interviews use standardised rubrics so every candidate is assessed against the same criteria, reducing the advantage that goes to candidates who “feel like a good fit” because they share the interviewer’s background. Skills-based assessment values what candidates can do over where they learned it; a bootcamp graduate who can build multi-agent systems is more valuable than an Oxbridge graduate who cannot.

Remote-first working is not just convenient; it enables geographic diversity. Talented developers in Dundee, Inverness, or Dumfries should not need to relocate to Edinburgh or London to work on cutting-edge AI. Flexible working supports people with caring

responsibilities, disabilities, or other circumstances that make traditional office hours difficult. Job descriptions are reviewed for biased language that discourages qualified candidates from applying.

11.5.2 AI for Scotland's Future: Schools Internship Programme

The most effective intervention for socioeconomic diversity is the earliest one. By the time talented young people from deprived areas reach the job market, they have already been filtered out by educational inequalities, lack of networks, and absence of role models. A schools internship programme addresses this directly.

The programme would offer paid summer internships for S5 and S6 students from schools in Scotland's most deprived areas, defined as the bottom 20% on the Scottish Index of Multiple Deprivation. These include former industrial communities in North Lanarkshire, South Ayrshire, Fife, and Dundee, as well as rural areas with limited technology exposure in the Highlands and Borders.

The placements last six to eight weeks and involve real project work on actual product features—not makework or observation. Interns are paid the real living wage, not offered unpaid “work experience” that only students with family financial support can accept. Each intern is paired one-to-one with a team member who provides mentorship throughout the placement.

High-performing interns are offered pathways to continue: university sponsorship (fees and living costs) for those pursuing higher education, or apprenticeships for those who prefer to enter the workforce directly. This creates genuine social mobility, not just a nice line on the company's website.

The programme scales with the company. In year one, a pilot with two interns from two schools tests the model and identifies improvements. In year two, six interns from six schools allow broader reach while maintaining quality mentorship. By year three, twelve interns from twelve schools makes this a significant annual commitment with formal school partnerships.

This is not corporate social responsibility theatre. I know from personal experience that talent is distributed everywhere but opportunity is not. Government policy shapes those opportunities—and so can technology companies that choose to act rather than simply talk about values.

Conclusion

Picture Sarah a year from now. She's completing her housing retrofit impact assessment—not in weeks of fragmented effort across disconnected templates, but with all eight assessment types working together for the first time. Every legal trigger identified. Every relevant piece of evidence found and available to reuse across future assessments. Every cross-assessment conflict surfaced for ministerial decision. The assessment that reaches ministers is thorough, consistent, and defensible. That's what this system delivers.

Better assessments, not just faster ones. Human judgement amplified by AI that finds more evidence, surfaces hidden conflicts, and catches what manual processes miss—but never makes decisions that belong to policy professionals. Seven phases of human checkpoints ensure no evidence enters any assessment without explicit approval. The architecture is production-proven: the same patterns power systems at Klarna and LinkedIn, and GraphRAG delivers 3–4× accuracy improvement over traditional retrieval.

The alternative is what's happening now: policy professionals using ChatGPT in the shadows, without oversight, without audit trails, with documented risks of hallucination. The choice is not whether AI will be part of impact assessment. The choice is whether it will be systematic, transparent, and accountable.

The design foundations are complete: fifteen personas, one hundred forty-seven user stories, approximately 70,000 words of wireframe specifications. Technical feasibility is validated. Fifteen weeks of co-design with policy professionals will transform these specifications into working software—and by Week 15, Scottish Government will have validated evidence that agentic AI can improve impact assessment quality, with a platform ready for wider deployment.

Sarah deserves tools that match the importance of her work. **I'm ready to build them.**