Ecaveo.

RESPONSIBLE AI
WHITE PAPER SERIES

PUBLIC SERVICES · CRITICAL NATIONAL INFRASTRUCTURE AND GOVERNMENT

The Known Safe State

Artificial intelligence across essential services, and the condition every deployment must be able to return to

Paul Forrest
Fractional Head of AI, Ecaveo
September 2026

About this paper

This paper covers government and critical national infrastructure, and it sits alongside a trilogy on professional firms, a paper on freight and warehouse operations, and one on financial services. Each stands alone.

The title is borrowed from control engineering, where a safe state is the condition a system falls back to when it can no longer be trusted to run. Public services need the same idea and mostly do not have it written down. If a model is withdrawn on a Tuesday morning, what does the service do on Tuesday afternoon, and has anybody rehearsed it? That question turns out to be a better organising principle for public sector AI than any of the ethics frameworks, and it is the one this paper is built around.

Editorial balance. Roughly seven parts in ten of this document is evidence and sector analysis, two parts in ten is original framework, and one part in ten concerns how I would run the work. The evidence on measured benefit is weaker than the volume of announcement suggests, and Chapter 2 says so at length.

HOW TO READ THE EVIDENCE TAGS

Every substantive claim carries a tag stating what kind of evidence supports it.

REGULATOR A regulator, government department, audit body or official statistics publisher speaking in its own voice.

STANDARD A published standard, statute, statutory instrument or formal guidance document.

CASE LAW A judgment, tribunal decision, public inquiry or enforcement action.

PEER-REVIEWED A primary study, systematic review or parliamentary investigation working from primary material.

SURVEY A survey or dataset with disclosed method and sample.

VENDOR RESEARCH Research from a body with a commercial or membership interest in the answer. Directional only.

PRACTICE Practitioner opinion, including my own.

EVIDENCE GAP A question the available evidence does not answer.

ECAVEO My own framework, proposed rather than proven.

Legal, regulatory and safety notice. This paper describes legislation, public law duties, regulatory guidance and safety obligations throughout, and Chapters 3 and 5 in particular. Nothing in this document constitutes legal advice, and nothing in it substitutes for a competent person's risk assessment or for a departmental legal opinion. The provision of legal advice sits outside the terms of any engagement with the author or with Ecaveo, and the material is presented to support discussion and further review by qualified advisers.

Statements about what a source says are attributed to that source. Where I draw a practical implication the source does not itself state, the text says so. Several of the instruments described here were still in passage or in draft when this was written.

Contents

THE ARGUMENT IN ONE PARAGRAPH

Government has adopted artificial intelligence at scale ahead of any evidence that it works, and the transparency machinery that is supposed to supply public legitimacy is a policy commitment rather than a legal duty, with 152 published records across a state that runs thousands of services. Critical infrastructure adds a harder problem, because a service that cannot go offline needs a defined condition to fall back to and a workforce that has practised operating in it. This paper argues that the safe state is the deliverable. Write down what the service does when the model is withdrawn, rehearse it, measure how long it takes, and the rest of the assurance apparatus has something real to attach to. Skip it, and every other control is a document.

ILLUSTRATIVE MATERIAL

Every authority, operator, caseworker and member of the public in the worked examples is invented, and each is marked as an illustrative scenario where it appears. Published research, official statistics, inquiry findings and enforcement decisions are cited and numbered.

CITATION

Forrest, P. (2026). *The Known Safe State: Artificial Intelligence in Critical National Infrastructure and Government*. Ecaveo Responsible AI Series, Public Services. Version 1.0, September 2026.

ALSO IN THIS SERIES

Partial Understanding. Financial services.

The Protected Constraint. Operations.

Verified Authority. Law firms.

The Traceable Thesis. Private equity.

A NOTE FROM THE AUTHOR

On the difference between a pilot and a service

Most of what has been published about AI in government describes pilots. Almost none of it describes what happens when a pilot becomes something a member of the public depends on.

I came to this from business process reengineering in the 1990s, then machine learning from 2014, and I build small specific models in preference to large general ones. Public sector work has been a recurring part of that, usually arriving through a portfolio company selling into government rather than through a department directly, which gives an odd but useful vantage point.

What I did not expect, when I went through the published material for this paper, was how thin the evidence of benefit is. The announcements are numerous and the evaluations are few, and of the evaluations that exist, none is a randomised controlled trial with an objective outcome measure. The cross-government Copilot experiment had 20,000 participants and no control group.[18] REGULATOR HMRC did randomly allocate licences to 3,000 employees, which is real methodological effort, and then measured self-reported time saving.[19] REGULATOR The most careful piece of work in the whole set measures accuracy on consultation analysis and reports no cost saving at all.[20] REGULATOR

None of that means the technology does not work. It means nobody has shown that it does, in this setting, on a measure that would survive scrutiny. That is a normal position for a technology at this stage and it is only a problem because of how the adoption is being described.

The other thing that struck me is the mismatch between the ethical apparatus and the operational question. There are principles, frameworks, toolkits and a transparency standard. There is very little about what a service does at 11am on a Tuesday when the model has been withdrawn and the queue is still there. That question is the whole of this paper.

TWO PAGES THAT CIRCULATE ON THEIR OWN

Executive summary

Adoption is broad, evidence of benefit is absent, transparency is a policy promise, and the cyber assurance framework that covers critical infrastructure has nothing in it about AI at all.

1

No UK government AI evaluation measures an objective outcome against a control. The cross-government Microsoft 365 Copilot experiment ran with 20,000 participants across 12 organisations and no control group, reporting an average 26 minutes a day saved on a self-reported basis, with 17 per cent reporting no clear time saving.[18] REGULATOR HMRC's phase three trial randomly allocated licences to 3,000 employees and still measured self-reported time.[19] REGULATOR The strongest study measures accuracy on consultation analysis and gives no cost figure.[20] REGULATOR

2

Algorithmic transparency is mandatory as policy, and it is not a legal duty. The Algorithmic Transparency Recording Standard binds central government departments and in-scope arm's length bodies through cross-government policy announced in February 2024 and formalised in December 2024, enforced administratively through digital spend controls. No statute or statutory instrument sits behind it.[6] STANDARD The published record count stood at 152 on 12 September 2026.

3

The Playbook does not require an inventory of live AI systems. The AI Playbook for the UK Government recommends a use cases register, described as a way of capturing use cases and prioritising which to explore first.[4] STANDARD That is a pipeline tool. Anybody who believes there is a government-wide duty to inventory deployed AI is describing something that does not exist.

4

The cyber assurance framework for essential services contains no AI content. Version 4.0 of the National Cyber Security Centre's Cyber Assessment Framework, released 18 April 2024 and reviewed 6 August 2025, has four objectives and 14 principles, none of them AI-specific.[10] STANDARD It is the framework used by competent authorities under the network and information systems regime and across the public sector through GovAssure.

5

Incident visibility is structurally incomplete. Five cyber attacks on UK water systems since 1 January 2024 were disclosed by the Drinking Water Inspectorate under freedom of information, and had not been public because the network and information systems regime requires reporting only where an attack causes disruption.[34] REGULATOR These did not.

6

Condition data was already being collected, and not acted on. The review into the North Hyde substation fire that closed Heathrow for most of 21 March 2025 found an elevated moisture reading in oil samples taken in July 2018, with mitigating actions appropriate to its severity not implemented.[35] REGULATOR The constraint there was organisational. More analysis would not have found it any earlier.

What follows for an accounting officer

Write down the safe state before buying the model. For every service where AI will influence a decision or an operation, record what the service does without it, who is competent to run it that way, how long the fallback can be sustained, and when it was last exercised. This is the one artefact that survives every vendor, every model change and every reorganisation.

Publish the transparency record early and treat the exemptions with suspicion. The standard permits withholding on freedom of information grounds including commercial sensitivity and risk of gaming, which is a broad carve-out and the obvious place for a department to hide.[6] STANDARD A record that omits the thing a citizen would want to know satisfies the policy and defeats its purpose.

Instrument the equality duty at the point of deployment. The Court of Appeal held in 2020 that a public authority breached section 149 by never having sought to satisfy itself that a software program did not have an unacceptable bias on grounds of race or sex, and it did so without finding that the software was in fact biased.[23] CASE LAW The duty is one of enquiry. A vendor's assurance does not discharge it.

Chapter 9 sets out an 18-month sequence. Chapter 10 states what would change the position taken here.

PART ONE · THE POSITION

A state that has adopted quickly and measured almost nothing

CHAPTER ONE

The service is the system

The unit of design in public services is never the model. It is the whole service, including the people, the operational technology, the suppliers, the manual fallback and the route a citizen takes when the answer is wrong.

Government and critical national infrastructure share a property that separates them from every commercial sector in this series. The service cannot simply stop. A bank can suspend a product line. A water undertaking cannot suspend supply, and a benefits agency cannot suspend payment, and the consequence of that is that every design decision has to carry an answer to the question of what happens when the new component fails.

The National Protective Security Authority defines critical national infrastructure as those critical elements of infrastructure, meaning assets, facilities, systems, networks or processes and the essential workers who operate them, whose loss or compromise could have a major detrimental effect on the availability, integrity or delivery of essential services, or a significant effect on national security, national defence or the functioning of the state.[31] STANDARD The list runs to 13 sectors, being chemicals, civil nuclear, communications, defence, emergency services, energy, finance, food, government, health, space, transport and water.

Those 13 sectors are not the same population as the regime that regulates cyber security for essential services. Schedule 2 to the Network and Information Systems Regulations 2018 covers 10 subsectors, being electricity, oil, gas, air transport, water transport, rail transport, road transport, healthcare, drinking water supply and distribution, and digital infrastructure, together with relevant digital service providers.[29] STANDARD Chemicals, civil nuclear, defence, emergency services, finance, food, government and space are therefore critical national infrastructure that mostly sits outside the essential-services duties. Anyone reasoning about UK CNI regulation from the network and information systems regime alone is reasoning about two thirds of it.

A MISMATCH WORTH KNOWING

The 13 CNI sectors and the 10 essential-service subsectors in the network and information systems regime are different lists. Eight CNI sectors sit largely outside that regime.

1.1 Interdependence, in the government's own words

The UK Government Resilience Framework of December 2022 is direct about the shape of the risk. The UK's critical national infrastructure is described as an interconnected system, whose interconnectedness brings many benefits and comes with risks, especially the possibility of cascading failures across systems, with interconnected vulnerabilities that may be significantly underestimated and issues that could be far reaching.[7] STANDARD

The Resilience Action Plan of July 2025 restates it and adds a specific example, noting that critical infrastructure failure has the potential to cause cascading and catastrophic consequences, including power outages affecting other essential functions, and that data centres may be dependent on a single power source.[8] STANDARD The 2026 Implementation Report, published on 14 July 2026, describes the United Kingdom as facing a volatile, diverse and interconnected risk landscape, and commits to a CNI Cyber Resilience Index piloting during 2026.[8] STANDARD

One caution about these documents, since they are frequently cited as though they were AI strategy. They are not. Artificial intelligence appears barely at all in the Action Plan, surfacing in a local resilience forum trailblazer exploring how AI could make communities more resilient, and as a driver of evolving biological threats.[8] STANDARD The 2026 report goes fur-

ther, identifying AI in its risk landscape section as presenting large, overlapping risks to the economy, jobs and social trust, including highly realistic deepfakes. Still, a reader looking for a government position on AI in critical infrastructure will not find one in the resilience documents.

1.2 The case that shows what the real constraint is

Late on the night of 20 March 2025 a high-voltage bushing failed on a transformer at the 275kV North Hyde substation in west London, most likely because moisture had entered the bushing and caused an electrical fault. The transformer caught fire.[35] REGULATOR

The consequences ran outwards in the way the resilience framework describes. Some 71,655 domestic and commercial customers lost power. Hillingdon Hospital, road and rail services and three data centres were affected. Heathrow lost one of its three independent supply points and closed for most of 21 March, reopening for repositioning flights late that day and returning to full operation on 22 March.[35] REGULATOR

The finding that matters for this paper sits earlier in the timeline. An elevated moisture reading was detected in oil samples taken in July 2018, nearly seven years before the fire, and mitigating actions appropriate to its severity were not implemented.[35] REGULATOR

Read that against the standard case for AI-driven predictive maintenance in infrastructure, which is that better analysis of condition data will find failures earlier. Here the data existed, the analysis had been done, the reading was elevated and somebody had written it down. What failed was the organisational path between a measurement and an intervention. My own reading, which the review does not state in these terms, is that most infrastructure operators would find the same thing if they looked, and that the marginal return on better detection is low in an organisation that does not act on the detection it already has. PRACTICE

Field inspection. Condition data has been collected in this sector for decades. The constraint on acting upon it has rarely been analytical.

1.3 What a safe state actually is

Borrowed from process safety, a safe state is the condition a system moves to when it cannot be trusted to continue. In an industrial plant it is usually well defined, physically enforced and tested. In a public service that has just introduced a model, it is usually undefined, which means it will be improvised on the day.

The four questions that constitute it are unglamorous and can be answered on one page. What does this service do if the model is withdrawn this morning? Who is competent to run it that way, and how many of them are there? For how long can that be sustained before something else breaks? When was it last exercised, with real staff and real volumes?

I would treat a service that cannot answer all four as not yet ready to deploy, whatever the model's accuracy. That is my own position and it is stricter than anything in the AI Playbook or the Data and AI Ethics Framework, neither of which requires a fallback to be defined or exercised.[4,5] STANDARD ECAVEO

CHAPTER TWO

Adopted ahead of the evidence

This is the most uncomfortable chapter in the paper and the most useful. Every published UK government evaluation is weaker than its coverage suggests, and the strongest one reports no saving.

Start with the audited baseline, because it is the only survey of government AI use conducted by an independent body with a statutory remit. The National Audit Office surveyed 89 government bodies in autumn 2023, achieving an unusually high 98 per cent response rate from 87 organisations, and published in March 2024.[13] REGULATOR

It found 37 per cent of bodies, being 32 of 87, had deployed AI, reporting 74 use cases in total, which is typically one or two each. Some 70 per cent were piloting or planning, with 61 bodies reporting a median of four use cases. A quarter had no AI activity at all. The barriers reported were skills recruitment and retention at 70 per cent, lack of funding also at 70 per cent, unclear liability at 67 per cent, limited access to good-quality data at 62 per cent, and concern about unreliable outputs at 57 per cent.[13] REGULATOR

The ATRS finding in the same report is the one that has travelled least and matters most. Of the 32 organisations with deployed AI, only eight reported being always or usually compliant with the Algorithmic Transparency Recording Standard.[13] REGULATOR

I could find no successor National Audit Office value-for-money report on AI in government between March 2024 and the date of this paper. The Public Accounts Committee's report of 26 March 2025 is the most recent parliamentary scrutiny, and it found that the standard was not widely used, that only 33 records had been published at January 2025, and that progress on embedding transparency had been slow.[14] PEER-REVIEWED

2.1 Three evaluations, read properly

The cross-government Copilot experiment. It ran with 20,000 participants across 12 organisations including the Department for Work and Pensions, HM Revenue and Customs, the Home Office and the Ministry of Justice, running from 30 September to 31 December 2024. There was no control group and no randomisation. Evidence came from Microsoft usage dashboards covering 14,500 users, a survey with 7,115 responses, feedback forms and five focus groups.[18] REGULATOR

The headline is an average of 26 minutes a day saved. That figure is self-reported and derived by taking the middle point of each categorical range offered to respondents. The report's own limitations are candid, noting that it was not possible to identify how time saved was spent, that the experiment was marketed as a finite period which may have inflated adoption, that user experience was not consistent, and that the final month was disrupted by the festive period. Some 17 per cent reported no clear time saving.[18] REGULATOR The document hedges appropriately. Its coverage did not.

The HMRC phase three trial. Licences randomly allocated to 3,000 employees plus around 500 volunteers, September to December 2024, with 1,364 survey responses at a 45 per cent response rate, Viva Insights usage data, focus groups and a 49-participant task-based exercise.[19] REGULATOR Random allocation is genuine methodological effort and it deserves credit. The outcome measure is still self-reported time saving, at 2 to 3 per cent of a working week, which HMRC extrapolates to £50m of annual productivity benefit at 50,000 licences.

The report describes the figure as a baseline estimate, and discounts the time savings by 20 per cent for non-usage variability. Security and data privacy concerns were cited by 46 per cent of non-users.

The Consult evaluation. This is the most careful piece of public sector AI evaluation I found in the UK. The incubator for artificial intelligence tested its consultation analysis tool on a live Scottish Government consultation on the regulation of non-surgical cosmetic procedures, across 5,354 question responses over six questions.[20] REGULATOR Exact-match agreement with human reviewers was 60 per cent, with an F1 score of 0.76 accounting for partial matches. Reviewers added themes 1,671 times and removed them 763 times, used the "other" category 793 times against the tool's 47, and reviewed at a median of 23 seconds per response with 77 per cent reviewed in under a minute.

THE COMPARATOR THAT CHANGES THE READING

Human reviewers agreed with each other on consultation theme coding only 62 per cent of the time. Against that benchmark, 60 per cent machine agreement is a good result.

The evaluation states its own weakness clearly, which is that the 62 per cent human agreement benchmark was measured in internal testing rather than on the same data, and that future evaluations will use double human review on identical material. It gives no cost saving figure at all. Headline savings claims circulating in press coverage are not supported by this document.

2.2 The exposure estimate that is not a productivity figure

The Alan Turing Institute, working with the Office for National Statistics and the Department for Transport, published an analysis in June 2025 using the ONS Public Sector Time Use Survey, drawing on 1,513 respondents and 3,005 diary entries weighted to population estimates.[21] PEER-REVIEWED Two researchers independently coded 80 of 91 work activities against a five-category exposure rubric, achieving 82 per cent direct inter-coder agreement.

The result quoted everywhere is that 41 per cent of an average public sector worker's 8.3-hour day, around 3.4 hours, is spent on activities that could be supported by generative AI, ranging from 33 per cent in healthcare to 49 per cent in primary and secondary education.[21] PEER-REVIEWED

The authors are careful in ways that most citations are not. They note that the coders do not directly work in public sector roles, that exposure scores were represented by their upper bound and may over-estimate exposure, that activities were assessed in isolation ignoring workflow interdependencies, and that the analysis excludes commercial costs, data security and privacy compliance, ecological and legal impacts, and worker trust. Only four activities, being 5 per cent, received the highest exposure classification.

The single most useful line in the report is the authors' own, which is that technical feasibility does not equal successful adoption. They cite a forthcoming study finding that only 4 per cent of teachers actually use generative AI for marking, despite moderate exposure scores.[21] PEER-REVIEWED A paper that quotes 41 per cent as a productivity opportunity without that qualification will be taken apart by anybody who has read the source.

WHAT UK GOVERNMENT HAS ACTUALLY MEASURED

STUDY	DESIGN	RESULT	WEAKNESS
GDS Copilot experiment, Jun 2025	20,000 users, 12 bodies, no control, not randomised	26 minutes a day, self-reported; 17 per cent saw no saving	No counterfactual; categorical estimates; time use unmeasured
HMRC phase 3, Jul 2026	Licences randomly allocated to 3,000 staff	2 to 3 per cent of a working week; £50m extrapolated	Outcome still self-reported; no formal control comparison
i.AI Consult, May 2025	Live consultation, 5,354 responses, human benchmark	60 per cent exact agreement, F1 0.76, against 62 per cent human agreement	Human benchmark measured on different data; no cost figure
Turing and ONS, Jun 2025	Expert coding of time use diaries, 1,513 respondents	41 per cent of the day is exposed to generative AI support	A theoretical upper bound, not a measured effect

TABLE 1 Four evaluations, graded by design. None is a randomised controlled trial with an objective outcome measure, and the strongest measures accuracy while reporting no saving.

2.3 The claim that needs a footnote it does not have

HM Revenue and Customs stated in its 2025-26 annual report that £10bn of tax was protected and recovered through artificial intelligence and advanced analytics.[42] REGULATOR That number will appear on slides for years.

Treat it with care. Artificial intelligence and advanced analytics is an undifferentiated category, no methodology is published in the executive summary, and no counterfactual is stated, meaning there is no estimate of what would have been recovered without it. It is a departmental claim and not an evaluated finding. EVIDENCE GAP I am not suggesting it is wrong. I am saying that nobody outside HMRC can currently tell whether it is right, and that a board being asked to fund something on the strength of it should ask for the derivation.

2.4 Scale, for context

The state this applies to is large. Civil service headcount stood at 557,825 at 31 March 2026, with 523,795 full-time equivalents, and the five largest departments account for 67.8 per cent of the workforce.[41] REGULATOR HMRC alone serves around 40 million individuals and 5.7 million businesses, collected £966.4bn in tax revenue, and reports 78 per cent of customer interactions as digital.[42] REGULATOR

The Department for Work and Pensions handled 43.2 million calls in 2023-24 across in-house and outsourced lines. Of the 26.2 million in-house calls, 3.6 million, being 14 per cent, were blocked or deflected before reaching a queue, and 17.3 million were answered. That is 76 per cent of the calls that reached a queue and 66 per cent of the calls received, against an 85 per cent target, with an average in-house wait of 15 minutes and 23 seconds.[43] REGULATOR Some 22.7 million people were receiving a benefit or pension payment in May 2023, with a projected caseload of 32.5 million by 2025-26.

Those volumes are why the opportunity is real and why the failure modes are serious. A 1 per cent error rate on a benefits decision process is a different object from a 1 per cent error rate on a marketing email.

The estate underneath them is in poorer condition than the adoption announcements imply. The government's own review of the state of digital government, published in January 2025, identified more than 140 AI use cases across central government, put roughly 28 per cent

of central government technology estates in the legacy category in 2024, recorded 22 per cent of red-rated legacy systems as lacking adequate remediation funding, and reported a digital recruitment campaign failure rate of 50 per cent in 2024 against 22 per cent in 2019.[17] REGULATOR Its headline claim of more than £45bn a year in unrealised savings from full digitisation is a potential estimate and not a measured benefit, and should be labelled as such wherever it is quoted.

CHAPTER THREE

Legitimacy, lawfulness and recourse

Public bodies are held to duties that no commercial firm faces. Three of them bite directly on automated decision-making, and one has already been litigated to the Court of Appeal.

3.1 The equality duty is a duty of enquiry

In August 2020 the Court of Appeal decided *R (Bridges) v Chief Constable of South Wales Police*. The appeal succeeded on three of five grounds, covering the sufficiency of the legal framework under Article 8(2), the adequacy of the data protection impact assessment, and the public sector equality duty.[23] CASE LAW Automated summaries of this judgment frequently get the count wrong, so it is worth stating.

The equality holding is the one with the widest application. The Court found that South Wales Police had breached section 149 of the Equality Act 2010 because it had never sought to satisfy itself that the software program did not have an unacceptable bias on the grounds of race or sex, and had failed to take reasonable steps to enquire whether the facial recognition software was biased.[23] CASE LAW

What makes the case so useful is what the Court did not find. There was no clear evidence before it that the software was in fact biased. The breach was procedural, because the duty under section 149 is one of enquiry and process, whose purpose the Court described as ensuring that public authorities give thought to whether a policy will have a discriminatory potential impact.

THE POINT MOST OFTEN MISSED

The Court found no clear evidence that the software actually was biased. The breach was in not having asked.

Two consequences follow for any public body deploying a model, and I would state both as reasonably firm rather than certain, since a different tribunal on different facts could reach a narrower view. A vendor's assurance that a system is unbiased does not discharge the duty, because the duty is on the authority to enquire. And commercial confidentiality over training data does not excuse the failure to enquire, though it may well shape what enquiry is possible.

3.2 The transparency standard, and what it is not

The Algorithmic Transparency Recording Standard is the machinery through which government intends to earn public legitimacy for algorithmic decision-making. Its scope was formalised in a Government Digital Service policy published on 17 December 2024, after an announcement in February 2024 and a rollout beginning that March.[6] STANDARD

The mandatory requirement applies to central government, meaning all ministerial departments, all non-ministerial departments, and those arm's length bodies that provide public or frontline services or routinely interact with the general public. Local government, police forces and devolved bodies publish voluntarily.[6] STANDARD

The standard is policy, and it is not law. No statute and no statutory instrument sits behind the ATRS. It is cross-government policy direction, and enforcement is administrative, running through the digital and technology spend controls to which an algorithmic transparency check was added in February 2025.[6] STANDARD Describing it as a legal requirement is a material error,

> and one a specialist reader will notice. That does not make it weak. Spend controls are a real lever, since the Playbook records that assurance is mandatory for all digital and technology spend above £100,000 for anything public facing and £1m for anything else.[4] STANDARD

Exemptions track the Freedom of Information Act 2000, and include national security, commercial sensitivity under section 43, risk of gaming or evasion of the tool, and infringement of intellectual property.[6] STANDARD Anybody who has drafted an FOI refusal will recognise the width of that. Commercial sensitivity in particular covers most of what a vendor would prefer not to publish.

The record count has grown and remains small against the size of the state. The Ada Lovelace Institute counted nine records in summer 2024. The Public Accounts Committee recorded 33 at January 2025. The Institute counted 56 in early March 2025.[14,16] PEER-REVIEWED The official finder returned 152 records across 38 organisations when I checked it on 12 September 2026.[6] STANDARD

Growth of that shape is real progress and it should be acknowledged. It also has to be read against the government's own commitment, recorded in Treasury Minutes of June 2025, which was to publish records for all currently identified in-scope algorithmic tools as at March 2025 in government departments, by the end of 2025, explicitly excluding arm's length bodies and anything identified after March 2025.[15] REGULATOR Two exclusions are built into the promise.

3.3 Automated decisions after February 2026

Section 80 of the Data (Use and Access) Act 2025 replaced Article 22 of the UK GDPR with Articles 22A to 22D, in force from 5 February 2026.[22] STANDARD For public bodies the change is significant and it is narrower than commentary sometimes suggests.

Article 22A defines a significant decision as one producing a legal effect or a similarly significant effect, and provides that a decision is based solely on automated processing where there is no meaningful human involvement in the taking of the decision. The general restriction on solely automated significant decisions is relaxed, with the full range of Article 6 lawful bases now available where special category data is not involved. Article 22B preserves the old near-prohibition where special category data is involved. Article 22C sets four safeguards that apply in every case, requiring the controller to provide information about decisions, enable representations, enable human intervention to be obtained, and enable decisions to be contested.[22] STANDARD

Meaningful human involvement is not defined in the Act. Article 22D leaves it to Secretary of State regulations, which have not been made. The Information Commissioner consulted on draft guidance on automated decision-making and profiling between 31 March and 29 May 2026, and is under a statutory duty arising from secondary legislation to produce an AI code of practice, neither of which had appeared by the date of this paper.[1] REGULATOR The Commissioner has also named the Department for Work and Pensions as an early focus for direct engagement on public sector automated decision-making.

So the position for a department today is that it is applying a relaxed rule whose central term it is defining for itself, under a regulator that has said what it thinks and has not yet said it in a form anyone can rely on.

3.4 Four cautionary cases, and what each is actually authority for

Precision matters here because these cases are cited loosely and a well-briefed critic will know the difference.

The Home Office visa streaming tool. A traffic-light risk rating applied to every UK visa application, with a non-public list of nationalities automatically flagged, creating a feedback loop where historic refusal rates drove future risk scores. The challenge was brought by the Joint Council for the Welfare of Immigrants and Foxglove. The Home Office conceded by letter on 3 August 2020 and discontinued the tool from 7 August.[24] CASE LAW It is often described as the first successful judicial review of a UK government algorithm. It was not a judicial review outcome at all, because it settled before any hearing, so there is no judgment and no precedent. As a pre-hearing concession it remains evidentially powerful and it is not case law.

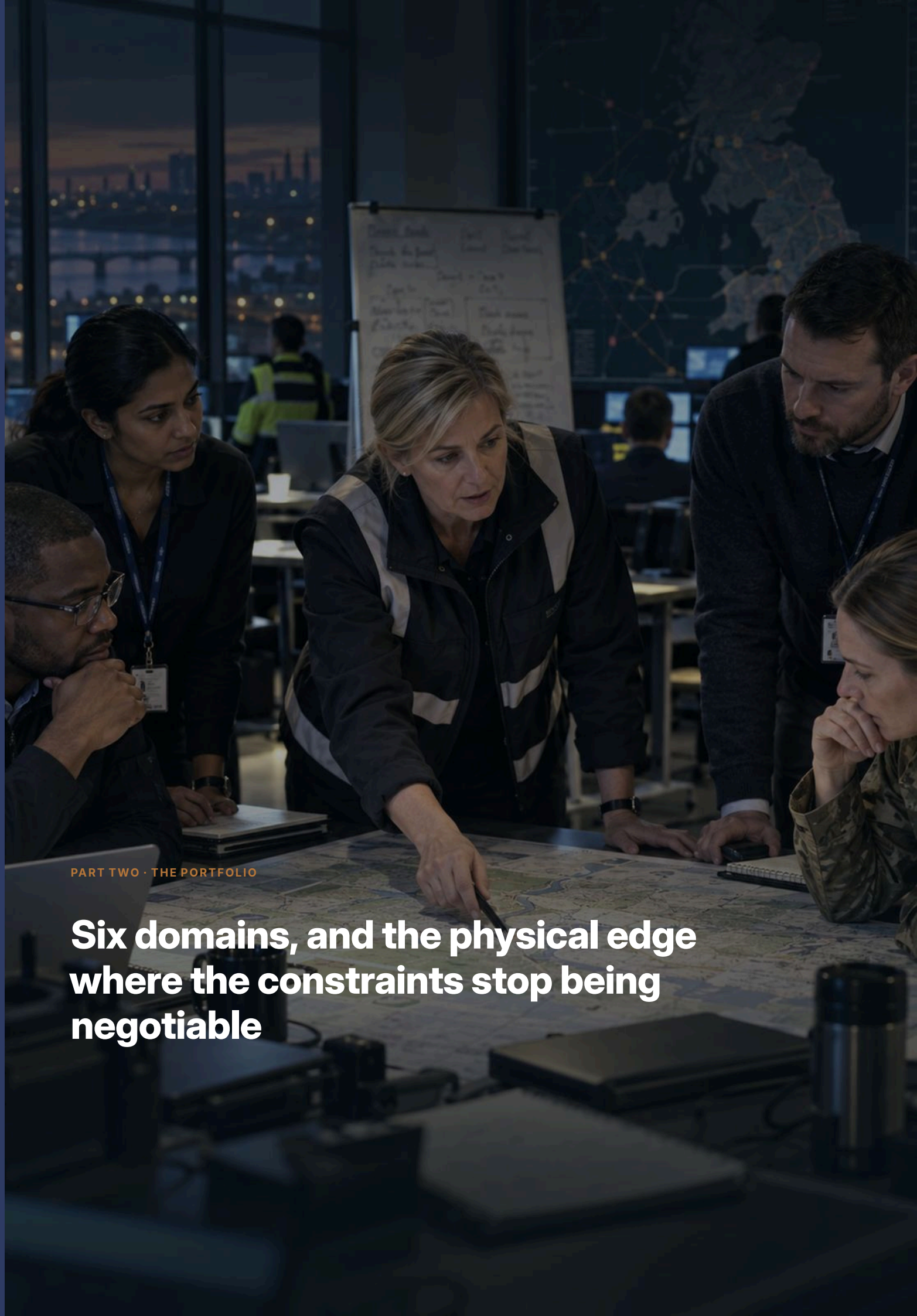
The 2020 qualifications algorithms. The Office for Statistics Regulation examined the models used by all four UK qualification regulators and found that all four failed to achieve public acceptance despite similar designs.[25] REGULATOR Its three principles are directly reusable as a governance frame, being to be open and trustworthy, to be rigorous and ensure quality throughout, and to meet the need and provide public value. Its central conclusion is that achieving public confidence in statistical models is not just about technical design. Note that the report does not state a downgrade percentage, and the widely quoted figure for grades downgraded is not sourced to it.

Post Office Horizon. Volume 1 of the inquiry report, chaired by Sir Wyn Williams, was presented on 8 July 2025. It records approximately 1,000 people prosecuted and convicted across the UK on Horizon evidence, a further 50 to 60 prosecuted but not convicted, around 10,000 eligible claimants in the redress schemes, 13 people who took their own lives in circumstances their families attribute to Horizon showing an illusory shortfall, at least 59 who contemplated suicide of whom 10 attempted it, and 19 who reported abusing alcohol as a consequence.[26] CASE LAW On causation the Chair is careful, writing that he cannot make a definitive finding of a causal connection between all 13 deaths and Horizon, while not ruling it out as a real possibility.

The relevance to AI is structural and I want to be careful about how I put it, because Volume 1 does not itself analyse the point. The scandal turned on a common-law presumption that computer systems are working correctly, which placed the evidential burden on individuals to disprove machine output they had no means to inspect. My own view, which goes beyond the inquiry's findings as published, is that opaque models in public administration reintroduce exactly that structure, and that the presumption is the thing to design against. PRACTICE

The Dutch cases. The District Court of The Hague held on 5 February 2020 that the SyRI welfare fraud detection legislation violated Article 8 of the European Convention on Human Rights and had no binding effect, on the grounds that the system was insufficiently transparent and verifiable, and that a state applying new technologies bears a special responsibility to balance benefits against privacy intrusion.[27] CASE LAW Separately, the childcare benefits scandal saw an initial 2017 report identify 232 families whose allowances were wrongly stopped, with more than 35,000 affected parents registered for compensation by April 2021, and the entire Dutch government resigning on 15 January 2021.[28] CASE LAW

The SyRI reasoning transfers because the United Kingdom remains a party to the Convention and Article 8 is incorporated through the Human Rights Act 1998. The reasoning that opacity can itself render an automated system disproportionate, independently of any demonstrated error, is available to a UK court. Whether one would accept it is a separate question and I have no basis for predicting it.



PART TWO · THE PORTFOLIO

Six domains, and the physical edge where the constraints stop being negotiable

CHAPTER FOUR

Where the value is credible

Six domains, ordered by how confidently the value can be demonstrated. The order differs from the one in most digital strategies, and the difference is instructive.

THE PUBLIC SERVICE PORTFOLIO

DOMAIN	OPPORTUNITY	VALUE TEST
Public services and casework	Summarise evidence, draft accessible correspondence, and route cases for human decision.	Timeliness, accuracy against a reviewed sample, reading level, cohort outcomes, appeals and complaint volume.
Incident detection and response	Correlate alerts, retrieve plans, assemble a timestamped situation report from approved sources.	Time to detection, time to a validated common picture, escalation quality and commander corrections.
Fraud, procurement and knowledge	Surface anomalies, retrieve policy, and strengthen commercial and programme insight.	Recovered value, false positive rate, decision quality and staff time released.
Demand and capacity	Forecast demand, queues, energy or resource requirements with stated uncertainty ranges.	Forecast error against the incumbent method, service level, reserve use and stability under shocks.
Asset health and inspection	Analyse sensor, image and maintenance evidence to direct engineering attention.	Detection lead time, false alarm rate, availability and maintenance burden. See the caution in section 4.2.
Emergency planning	Explore scenarios, identify dependencies and support cross-organisation exercises.	Dependency coverage, exercise findings, action closure and tested recovery time.

TABLE 2 Six domains. The value test states what would have to improve against a measured baseline. Where the evidence for a domain is weak, the entry says so.

4.1 Why casework sits at the top

Correspondence and casework has the strongest fit, and the Consult evaluation is the reason.[20] REGULATOR It demonstrates three things that transfer. There is a measurable ground truth, since a human decision exists to compare against. The human benchmark is lower than people assume, at 62 per cent inter-reviewer agreement on theme coding, which reframes what counts as acceptable machine performance. And the review step is cheap, at a median of 23 seconds per response.

Volumes make it worth doing. The Parliamentary and Health Service Ombudsman handled 123,987 contact centre enquiries in 2024-25 and accepted 38,045 complaints for consideration.[45] REGULATOR The Local Government and Social Care Ombudsman dealt with 22,010 complaints and enquiries, completed 4,441 detailed investigations, and upheld 83 per cent of them, having seen a 15 per cent increase in new complaints for the second consecutive year.[45] REGULATOR

That 83 per cent uphold rate is worth pausing on for a different reason. It suggests the complaints reaching that stage are largely well founded, which means the upstream decisions being complained about are going wrong at a rate the system has already priced in.

4.2 Why asset health sits lower than the market thinks

Predictive maintenance is the most heavily marketed AI application in infrastructure and it has the weakest published evidence base of anything in the table above. A PRISMA-compliant systematic review of digitalised predictive maintenance in railways, published in March 2026, identified 632 papers, screened 189 after deduplication and included 73 peer-reviewed articles from 2015 to 2026.[40] PEER-REVIEWED

Its findings are unflattering and they are specific. Most systems function only as unidirectional digital representations, so the bidirectional self-correcting digital twin largely does not exist in the field. Most of the literature is laboratory and simulation work, with deployment remaining predominantly theoretical. Deep learning models are described as challenging to interpret for infrastructure managers, and most digital twin applications rely on closed proprietary software ecosystems.[40] PEER-REVIEWED

Water offers the same picture from a different angle. Leakage across the 17 water and wastewater companies ran at approximately 2,869 million litres a day in 2024-25, around 20 per cent of all water put into supply, and only four of the 17 met their leakage targets.[44] REGULATOR Nobody would suggest the constraint there is a shortage of analysis.

The most interesting finding is structural. The review identifies what it calls a safety paradox, which is that railways replace components before failure, producing a severe deficiency of run-to-failure data for training models.[40] PEER-REVIEWED That is a permanent condition of safety-critical infrastructure and not a temporary data gap, and it will not be solved by collecting more of the data that is already being collected.

Put that alongside North Hyde from Chapter 1 and a picture emerges that ought to change procurement behaviour. The training data is structurally scarce, and the organisational route from detection to action is the demonstrated failure point. An operator that buys analytics before fixing the second will have bought a better version of a signal it already ignores.

4.3 Demand forecasting, which is quietly the best bet

Nobody writes strategy documents about demand forecasting, which may be why it works. The task has a ground truth that arrives on its own, since tomorrow's demand becomes known tomorrow. Backtesting against historical data is straightforward. The incumbent method is usually a spreadsheet with a seasonal adjustment, which is a beatable benchmark. And a forecast that is wrong is visibly wrong, quickly, without anybody being harmed.

If I were sequencing a first year of public sector AI work with a fixed budget and a requirement to produce defensible evidence, this is where I would start, and I would be prepared for it to be the least interesting item in the portfolio to everybody senior.

CHAPTER FIVE

The cyber-physical edge

Connecting an analytical system to an operational one expands the attack surface, and the framework that assures cyber security for essential services has nothing in it about artificial intelligence.

Operational technology changes the deployment question in a way that most AI governance material does not address. In an office environment the worst outcome of a compromised system is usually the loss or corruption of information. In an industrial environment it is a physical event.

The National Cyber Security Centre published Secure Connectivity Principles for Operational Technology on 14 January 2026, co-sealed by eight agencies across seven countries, covering Australia, Canada, Germany, the Netherlands, New Zealand, the United Kingdom and the United States.[9] STANDARD The guidance names the benefits of connectivity in operational technology environments as real-time analytics, predictive maintenance and remote monitoring and administration. It then records separately that the increasing use of third-party vendors, remote access solutions and supply chain integrations all expand the potential attack surface. Those two statements are the whole argument for treating an analytics deployment as a security change.

The eight principles are worth reproducing, because they translate directly into requirements a buyer can put in a specification.

- Balance the risks and opportunities, documenting a formal business case for each connectivity decision.
- Limit the exposure of connectivity, preferring outbound-only connections and brokered access through a demilitarised zone.
- Centralise and standardise network connections.
- Use standardised and secure protocols, migrating away from insecure industrial protocols.
- Harden the operational technology boundary with modern boundary devices, multi-factor authentication and least privilege.
- Limit the impact of compromise through segmentation, micro-segmentation and separation of duties.
- Ensure all connectivity is logged and monitored.
- Establish an isolation plan covering site, application and hardware-enforced isolation, and test it.

The last one is the safe state in cyber form, and it is the principle most often present in a document and absent from a rehearsal schedule.

5.1 The gap in the assurance framework

Version 4.0 of the Cyber Assessment Framework was released on 18 April 2024 and reviewed on 6 August 2025. It has four objectives and 14 principles, covering managing security risk, protecting against cyber attack, detecting cyber security events, and minimising the impact of incidents.[10] STANDARD It is used by competent authorities under the network and information systems regime, by CNI operators, and across the public sector through GovAssure.

There is no AI-specific objective, principle or contributing outcome anywhere in it. A CNI operator can be fully compliant with the framework its regulator uses while running models it cannot explain, on data it cannot trace, from a supplier it cannot substitute.[10] STANDARD That is not a criticism of the framework, which was written before the current wave and does a great deal of useful work. It is a statement of where the assurance boundary currently sits.

THE CONCRETE GAP

No AI-specific objective, principle or contributing outcome exists in the framework that assures cyber security for UK essential services.

The nearest thing to a filling of that gap is happening in the United States. On 6 April 2026 the National Institute of Standards and Technology published a concept note for a risk management framework profile on trustworthy AI in critical infrastructure, covering operators adopting AI across information technology, operational technology and industrial control systems, with the stated purpose of guiding operators towards specific risk management practices and helping them communicate trustworthiness requirements in an actionable way.[3] STANDARD It remains a concept note and not a published profile, and it carries no UK legal force. It transfers because UK operators and their suppliers routinely map to NIST frameworks alongside the Cyber Assessment Framework, and because no comparable UK work is in progress that I could find.

Note in passing that the generative AI profile itself, NIST AI 600-1, was published on 26 July 2024 and has not been revised. Citations to an April 2026 update are confusing it with the critical infrastructure concept note.[3] STANDARD

5.2 What incidents are actually visible

Five cyber attacks on UK water systems were reported to have occurred after 1 January 2024, disclosed by the Drinking Water Inspectorate in response to a freedom of information request and reported in November 2025.[34] REGULATOR They had not been public, and the reason is structural. The Network and Information Systems Regulations trigger reporting on service disruption, and none of these caused disruption.

That single fact tells a board more about the state of UK CNI incident visibility than any threat report. The published incident record is a record of disruptions, and intrusions that were caught, or that did not progress, or that were simply quiet, do not appear in it. Any resilience assessment built on the public record is therefore built on a filtered sample, and the filter selects out exactly the near-misses that would be most informative.

Where the record does exist it is instructive. In May 2026 the Information Commissioner fined South Staffordshire Plc and South Staffordshire Water Plc £963,900 following a breach affecting 633,887 individuals.[33] CASE LAW Initial access came through phishing in September 2020, with roughly 20 months of undetected dwell time before the main breach activity in mid-2022, discovered on 15 July 2022 during an investigation into IT performance. The Commissioner found limited controls allowing privilege escalation to administrator, monitoring covering only 5 per cent of the IT environment, obsolete unsupported software including Windows Server 2003, unpatched critical systems and no security scanning. The Commissioner's Interim Executive Director for Regulatory Supervision put it that waiting for performance issues or a ransom note to discover a breach is not acceptable, and that proactive security is a legal requirement and not an optional extra.

Be accurate about what this case is. The notice concerns personal data, and it does not state that operational technology or water supply was affected. It is an information security failure at a water undertaker with a two-year dwell time and 5 per cent monitoring coverage, which is damning without any embellishment.

5.3 The regime that is coming, and has not arrived

The Cyber Security and Resilience Bill was introduced on 12 November 2025 and had completed all Commons stages, reaching report stage in the House of Lords by early September 2026.[30] STANDARD It has not received Royal Assent. Anybody describing it as an Act is ahead of Parliament.

What it proposes matters for planning. New entities are brought into scope, including data centres, managed service providers, designated critical suppliers within essential-service supply chains, and large load controllers. Reporting moves to an initial notification within 24 hours of first awareness and a full notification within 72 hours. Penalties rise substantially, and the Secretary of State gains powers to issue strategic priority statements to regulators, to direct organisations and regulators to act for national security, and to update the regime by secondary legislation.[30] STANDARD Implementation is phased, and commentary suggests it might not be fully in force until 2028.

One criticism is worth carrying, because it bears on how much comfort the Bill should provide. Its sectoral scope excludes retail and automotive firms of the kind that suffered significant 2025 breaches, so an event on the scale of the retail ransomware wave described in Chapter 6 would still fall outside the regime.[30] STANDARD

CHAPTER SIX

Common-mode failure

The distinctive risk of AI in public services is not that one model fails. It is that many services fail at once because they depend on the same thing.

A common-mode failure is one where a single fault takes down components that were assumed to be independent. Public services are unusually exposed to it, for a reason that has nothing to do with technology, which is that they buy from the same small set of suppliers under the same frameworks and run them on the same infrastructure.

The clearest recent demonstration was not an AI event at all. On 19 July 2024 a faulty content update from a single security vendor affected 8.5 million Windows devices worldwide, according to Microsoft.[36] VENDOR RESEARCH A commercial cyber-risk quantification vendor modelled the UK cost at between £1.7bn and £2.3bn, built from endpoint detection market data, the vendor's market share, and an assumed average downtime of one working day across business interruption, response and post-response categories.[36] VENDOR RESEARCH That is a modelled estimate produced and funded by a firm that sells cyber-risk quantification, with no sector breakdown and no organisational sample disclosed, and its own report notes that 97 per cent of systems were fixed after nine days, which cuts against the 24-hour assumption in both directions. Use the number as an order of magnitude and nothing more.

The mechanism is the part to take seriously. Thousands of organisations that had never made a decision to depend on each other discovered that they did, through a component they had all separately chosen for good reasons.

The same shape appeared in the 2025 UK retail ransomware wave. The Cyber Monitoring Centre categorised the incidents affecting two major retailers as a single combined category two systemic event, estimating the financial impact at between £270m and £440m and publishing no central estimate.[32] PEER-REVIEWED Its method models impact from public and commercial transaction-level spending data, so it is an estimate and not an audited loss. The Centre excluded other retailers affected in the same period for want of information about cause and impact, which means the published range is a floor.

6.1 The cloud position, and the decision that did not happen

The Competition and Markets Authority opened its cloud services market investigation on 5 October 2023, published provisional findings on 28 January 2025, and issued its final report and decision on 31 July 2025.[38] REGULATOR It found that Amazon Web Services and Microsoft each hold between 30 and 40 per cent of UK infrastructure-as-a-service by value, and identified three adverse effects on competition, being market concentration and barriers to entry, technical and commercial barriers to switching including egress fees and limited interoperability, and Microsoft's software licensing practices. Committed spend agreements had been flagged in the provisional findings and were dropped from the final adverse findings, the Authority concluding that in their current form and application they do not harm competition in cloud services markets.
The inquiry group recommended that the Authority prioritise commencing strategic market status investigations into the two largest providers.

On 31 March 2026 the Authority declined to prioritise those investigations.[38] REGULATOR It opened a strategic market status investigation into Microsoft's business software ecosystem instead, launching in May 2026, and continued what it described as participative dialogue with both firms on cloud, committing to review progress in six months. Its stated reasons were that both firms had reduced egress fees, improved interoperability, extended EU Data Act standards to the UK and committed to formal processes for interoperability information requests. It prioritised business software partly because of the rapid integration of AI into workplace tools.

So the exit and portability problem in UK cloud is, as at September 2026, unresolved by regulation and addressed through voluntary commitments. Any statement that the Authority is investigating cloud strategic market status is out of date by six months.

6.2 Drawing the inference carefully

Four things are established. Two providers each hold between 30 and 40 per cent of UK infrastructure-as-a-service by value. The Authority found three adverse effects on competition and identified specific switching barriers. It then declined to use its strongest tool on cloud in March 2026. And government AI adoption is being delivered substantially through Microsoft 365 Copilot, across 12 organisations and 20,000 participants in the cross-government experiment alone.[18,38] REGULATOR

The inference is mine and I will state it as such. UK government AI adoption is deepening dependence on a market that its own competition authority has already found to be working badly, and it is doing so faster than the remedies are arriving. PRACTICE The Authority has made no such finding, and a reasonable person could argue that the voluntary commitments will hold. My concern is that the commitments were given by firms whose incentive to keep them declines as substitution becomes harder, and substitution is becoming harder rather than easier as models embed into workflow.

The procurement route is worth a line. The Crown Commercial Service operates RM6200, an Artificial Intelligence Dynamic Purchasing System, which lets suppliers join throughout its life.[39] STANDARD That is a materially better concentration profile than a closed framework. I could not obtain the current supplier count, the total spend through it, or the concentration of that spend, so I cannot say whether the better structure has produced a better outcome. EVIDENCE GAP

6.3 What a good recovery report looks like

The British Library's review of its own ransomware attack is the best-documented public body lessons report I have found in the UK, and it is worth reading in full by anybody responsible for a public service.[37] REGULATOR The attack occurred on Saturday 28 October 2023, attributed to the Rhysida group, with likely initial access through compromised privileged account credentials on a terminal services server. Around 600GB was exfiltrated, comprising just under half a million individual documents and 22 databases forcibly copied. Service was severely restricted for approximately two months and restoration continued well beyond.

The Library states plainly that it made no payment to the criminal actors and did not engage with them in any way. Its 16 lessons include enhancing network monitoring, fully implementing multi-factor authentication, implementing network segmentation, managing system lifecycles to eliminate legacy technology, prioritising remediation of legacy technology issues, prioritising recovery alongside security, ensuring cyber risk awareness at senior level, and proactively managing staff and user wellbeing.[37] REGULATOR

That last item does not appear in any framework I know of and it should. The recovery took months and it was carried out by people who had been through the incident, and treating their capacity as unlimited is the sort of planning assumption that only fails once.

ILLUSTRATIVE SCENARIO

Calderbrook Water rehearses the safe state

An invented water undertaking has run an anomaly detection model on pump telemetry for 14 months, and it has become the way the control room prioritises its morning checks. Nobody has written down what happens without it.

The exercise is scheduled for a Wednesday and lasts six hours. The model is withdrawn. Three things surface. Two of the four control room staff on shift have never worked without it, having joined after deployment. The manual prioritisation procedure exists as a document that references a screen layout replaced in 2023. And the daily check list, which the model implicitly generated, has to be reconstructed from memory by the shift supervisor, who leaves in March.

None of that is a technology failure. All of it would otherwise have been discovered during an incident, and the cost of finding it in an exercise was six hours and a facilitator.

PART THREE · THE WORK

Three deployments, the safe state, and an 18-month sequence

CHAPTER SEVEN

Three deployments, and how to evaluate them

Each is bounded, reversible and evaluable inside a quarter. None takes a control action or decides anything about a person.

7.1 The incident evidence assistant

Scope. Build a timestamped, source-linked operational picture from approved logs, reports and situation updates during an incident. The system takes no control action, issues no instruction and makes no recommendation about response options.

Why this one. During an incident the scarcest resource is a validated common picture, and assembling one currently consumes the attention of the people who should be deciding. The evaluation is unusually tractable because incidents are exercised routinely, which means a test environment already exists and nobody has to build one.

Evidence to collect. Event recall against the reconstructed record, meaning the proportion of material events the system captured. Source accuracy, meaning whether each entry resolves to a real record. Latency from event to appearance. Access control behaviour under multi-agency conditions, which is where these systems leak. Commander correction rate and correction type.

Stop condition. Any entry in a situation report that cannot be traced to a source. Any cross-organisation access leak.

7.2 The asset anomaly adviser

Scope. Shadow-mode analysis for one bounded asset class, with engineering disposition retained entirely by the engineer. No work order is raised by the system.

Why this one. Section 4.2 sets out why I rank this lower than the market does, and shadow mode is precisely how a sceptical operator should test the claim. The safety paradox means run-to-failure data is scarce, so the honest first question is whether the model can detect the failures the organisation already knows about, before anybody asks whether it can find new ones.[40] PEER-REVIEWED

Evidence to collect. Detection lead time against the incumbent method. Missed faults, established retrospectively against the maintenance record. Nuisance alert rate at the working threshold, which determines whether engineers will keep looking at it. Drift over the shadow period. Safe-state behaviour when input data goes stale or a sensor fails.

Stop condition. A nuisance rate that engineers stop responding to. That is a behavioural threshold rather than a statistical one, and it should be set by the engineers before the pilot starts.

7.3 The public correspondence assistant

Scope. Draft plain-language responses from approved policy, for caseworker review, with recourse routes stated in every response. The system does not send anything and does not decide anything.

Why this one. The Consult evaluation gives a method that transfers, the volumes are large enough for statistical power, and the review step is genuinely cheap.[20] REGULATOR The accessibility gain is real and under-discussed, since plain-language drafting is a skill that varies widely across a large caseworking population.

Evidence to collect. Accuracy against a reviewed sample. Reading level, measured and not assumed. Cohort outcomes, defined and instrumented before launch. Correction rate and type. Complaint volume against a matched baseline. Disclosure quality, meaning whether the recipient can tell that a machine drafted the response and what to do if it is wrong.

Stop condition. Any response reaching a member of the public that states a policy position not present in the approved source.

WHAT EVERY ONE OF THESE HAS IN COMMON

01
No action
The system produces evidence or a draft. A person decides. Nothing is sent, raised, closed or actuated by the machine.

02
Traceable
Every assertion resolves to a source record that a reviewer can open. Untraceable output is treated as a defect and not as a limitation.

03
Cohort measured
Outcome measurement by cohort is instrumented before launch, because it cannot be reconstructed afterwards.

04
Reversible
Withdrawal is a decision somebody can take on the day, against a written and exercised fallback.

FIGURE 1 Four properties, and the fourth is the one that gets cut.

Proposed design constraints for a first wave of public sector deployments. ECAVEO The traceability requirement follows the transparency and explanation expectations in the Data and AI Ethics Framework and the Article 22C safeguards.[5,22]

CHAPTER EIGHT

Writing down the safe state

One page per service, four questions, and an exercise in the diary. This is the framework the paper exists to argue for, and it is deliberately small.

Everything in the assurance apparatus attaches to something. Impact assessments attach to a purpose, validation attaches to a model, transparency records attach to a tool. The safe state attaches to the service, which is why it survives when the model, the vendor and the team have all changed.

I am proposing a single artefact, owned by the service owner, reviewed annually and after every material change. It is not a substitute for a data protection impact assessment, an equality analysis or a transparency record, all of which remain necessary. It sits underneath them and answers the question they do not ask.

THE SAFE STATE RECORD

FIELD	WHAT IT RECORDS	EVIDENCE THAT IT IS REAL
Fallback description	What the service does, concretely, from the moment the system is withdrawn. Named procedure, not a principle.	A document that references screens and forms that currently exist.
Competence	Who can run the service that way, how many of them there are, and where they sit.	A named list, with the proportion who have done it in the last 12 months.
Endurance	How long the fallback can be sustained before something else fails, and what fails first.	A stated figure with a reason, and the constraint that produces it.
Last exercised	The date, the conditions, the volume handled and what went wrong.	An exercise report, including the findings that were uncomfortable.
Withdrawal authority	Who can withdraw the system, on what evidence, without seeking approval elsewhere.	A name and a role, in a document that person has seen.
Reinstatement test	What has to be true before the system goes back on.	A written test, agreed before the first withdrawal rather than during one.

TABLE 3 One per service where AI influences a decision or an operation. Proposed framework. ECAVEO Nothing in the AI Playbook or the Data and AI Ethics Framework requires a fallback to be defined or exercised, which is the gap this fills.[4,5]

8.1 Why the exercise is the part that matters

A written fallback that has never been run is a hypothesis. The Calderbrook scenario in Chapter 6 is invented, and the failure modes in it are not, because they are the ones that recur. Staff who joined after deployment have never worked the manual way. Documentation references interfaces that have been replaced. Implicit knowledge that the system encoded has left the building with the person who encoded it.

None of those is detectable from a desk. All of them are detectable in half a day with a facilitator and a decision to actually turn the thing off.

The UK Resilience Academy is committed to training more than 4,000 people annually and the Crisis Management Excellence Programme covers more than 9,000 civil servants, so the exercise infrastructure exists.[8] STANDARD What I have not seen is an exercise programme that treats withdrawal of an analytical system as a scenario in its own right, and I would put one in the calendar for every tier three or tier four deployment.

8.2 Tiering, and why by consequence

Four tiers, classified by the consequence of an error and not by the sophistication of the system. A simple rules engine that stops a payment outranks a large model that drafts an internal summary, and any scheme that sorts the other way will spend its assurance effort in the wrong place.

RISK TIERS FOR PUBLIC SERVICE DEPLOYMENTS

TIER	TYPICAL SCOPE	MINIMUM CONTROLS
One	Public or non-sensitive internal content, reversible assistance	Approved tool, training, source checking and human review before anything is published.
Two	Official information or internal workflow decisions	Use-case record, data-flow review, evaluation against a baseline, logging and a named reviewer.
Three	Material affecting a member of the public, an operational asset, or a supplier relationship	Data protection impact assessment, equality analysis, transparency record, representative testing, cohort measurement and a written safe state.
Four	A decision about a person's rights or entitlement, a control action, or anything touching safety	Senior accountable owner, independent verification, strict permissions, continuous monitoring, exercised safe state, incident plan and rehearsed rollback.

TABLE 4 Consequence of error drives the tier. The controls are cumulative, so a tier four system carries everything from tiers one to three as well.

8.3 Six technical controls that apply at every tier

These are consistent with Information Commissioner accountability expectations, the National Cyber Security Centre's secure AI development guidelines of November 2023, co-sealed with 21 other international agencies, and the NIST generative AI profile of July 2024, which is a United States voluntary framework carrying no UK legal force.[1,2,3] STANDARD

- Enterprise identity and service permissions are enforced before retrieval and before any tool call, and not applied to results afterwards.
- Contracts define training use, retention, hosting region, subprocessors, deletion, incident notification, model change notification and audit rights. The Department for Work and Pensions supplier policy is a usable template, prohibiting the use of material classified OFFICIAL-SENSITIVE or above, personal data, and non-public departmental code in AI tools without explicit approval.[11] STANDARD
- Agents operate with least-privilege, short-lived credentials, and any consequential action requires explicit approval and is reversible.
- Every production system records the model, prompt, retrieval corpus, tools, permissions and evaluation version behind each output.
- Prompt injection, malicious documents, sensitive-data disclosure and insecure output handling are tested before scale.
- Generated work passes the same review and record-keeping controls as work produced without AI.

Organisations wanting a certifiable management system wrapped around the same ground have ISO/IEC 42001, published in December 2023.[46] STANDARD Certification is not a substitute for any of the six, and it is a reasonable way to make an internal programme legible to a supply chain.

CHAPTER NINE

From pilot to service, in 18 months

A sequence that produces evidence before it produces exposure, and that leaves an organisation defensible whatever the legislative position turns out to be.

9.1 The first 30 days

Interview leaders and observe real work across the priority functions. Watching beats asking, and the gap between what people describe and what they do is where the use cases are.

Inventory approved and shadow AI use, including capability embedded in software the organisation already licenses. The Playbook's use cases register is a pipeline tool and will not find what is already running.[4] STANDARD

Map data boundaries, statutory duties, permissions and the individuals who currently own each decision affecting the public. This is the input to both the equality analysis and the safe state record.

Publish interim safe-use guidance and a route for proposing an experiment. Organisations that skip this run the discovery exercise twice.

Select three pilots with baselines, representative test sets and stop conditions written down before anyone sees a tool.

9.2 The rest of the sequence

THE 18-MONTH SEQUENCE

PHASE	TIMING	OUTPUTS
Frame	Weeks 1 to 6	Service map, decision inventory, model and supplier register, outcome baselines, named accountability, and a draft safe state for each candidate service.
Prove	Weeks 7 to 16	Three shadow-mode pilots with cohort testing, held-out evaluation sets and independent challenge from outside the delivery team.
Assure	Months 5 to 9	Impact assessments, equality analyses, transparency records drafted and submitted, red-team findings closed, and the first live safe-state exercise completed with findings published internally.
Sustain	Months 10 to 18	Proven patterns extended, monitoring for outcomes, drift and supplier change, annual exercise scheduled, and manual competence maintained as a measured quantity.

TABLE 5 Each phase produces evidence the next depends on. The exercise in phase three is the item most often deferred and the one that most changes what a service actually does.

The Sustain phase contains the item that distinguishes a service from a project, which is manual competence maintained as a measured quantity. Once a service depends on a system, the proportion of staff who can operate without it begins to decay through ordinary turnover, and nobody notices because nothing goes wrong until something does.

9.3 What good looks like after 18 months

Scaled workflows with measured improvement against baselines set in month one. A single intake route, approved systems, documented use cases and monitored supplier changes. Transparency records published for every in-scope tool, including the ones that were awkward. Staff who can explain what a system cannot do and escalate a failure when they meet one. An exercised safe state for every tier three and tier four service, with a report that somebody senior found uncomfortable. And a backlog that separates proven opportunities from research and from uses the organisation has decided against.

The uncomfortable exercise report is the leading indicator. An exercise that goes well has usually been designed to go well.

CHAPTER TEN

Where the evidence is thin

The claims in this paper that would change if the evidence changed, stated plainly.

10.1 What I could not establish

There is very little randomised or quasi-experimental evidence on AI in UK public services and essentially none in critical infrastructure asset management. That is a finding and not a gap in my research. EVIDENCE GAP The systematic review of railway predictive maintenance found the literature dominated by laboratory and simulation work, with field-validated implementation largely absent.[40] PEER-REVIEWED

No official UK statistic exists for the combined economic size or employment of the 13 critical national infrastructure sectors. The National Protective Security Authority defines the sectors and publishes no economic aggregate, and I have not constructed one. EVIDENCE GAP

I could not obtain the current supplier count, total spend or spend concentration for the Crown Commercial Service artificial intelligence purchasing system, so the concentration argument in Chapter 6 rests on the cloud market evidence and not on procurement data. EVIDENCE GAP

The number of published transparency records moved between retrievals during the same session, returning 152 unfiltered and 138 on one sorted query. I have used 152 as the unfiltered total on 12 September 2026 and a reader should re-check it, since the number is the kind that gets quoted long after it stops being current.[6] STANDARD

Whether further volumes of the Post Office Horizon inquiry report have been published since the Chair's progress update of 8 July 2026, I could not confirm.[26] CASE LAW

10.2 What would change the argument

A randomised evaluation of a government AI deployment with an objective outcome measure. One properly designed study would do more for public sector decision-making than another two years of adoption announcements, and it would settle whether Chapter 2 is right.

An AI objective or set of principles added to the Cyber Assessment Framework. That would close the gap identified in Chapter 5 and would give CNI operators a regulatory hook they currently lack.[10] STANDARD

Royal Assent for the Cyber Security and Resilience Bill, and the commencement regulations that follow. Chapter 5 is written against a Bill in passage and would need rewriting against an Act.[30] STANDARD

Secretary of State regulations defining meaningful human involvement, or final Information Commissioner guidance. Chapter 3 turns on an undefined statutory term.[22] STANDARD

A statutory footing for algorithmic transparency. If the standard moved from policy to law, the argument in Chapter 3 about the exemptions would change shape, though it would not disappear, since the freedom of information exemptions would presumably come with it.

10.3 Where I have gone beyond the sources

The safe state framework in Chapter 8 is mine. Nothing in the AI Playbook, the Data and AI Ethics Framework, the Responsible AI Toolkit or the transparency standard requires a fallback to be defined, still less exercised.[4,5,12] STANDARD ECAVEO I think that is the most consequential omission in the current guidance and I accept that a reasonable person could regard it as an operational matter for each service to handle in its own way.

The reading of North Hyde in Chapter 1, that the constraint on infrastructure resilience is organisational and not analytical, is my inference from the review's findings. The review does not say it. PRACTICE

The claim in Chapter 6 that government AI adoption is deepening dependence on a market its own competition authority has found to be working badly is mine, and I have set out in that chapter both the evidence for it and the reasonable counter-argument. PRACTICE

One thing I have thought about and cannot resolve concerns the transparency exemptions. A department that publishes a record omitting the mechanism, on grounds of gaming risk, has complied with the standard and defeated its purpose, and the gaming risk is often perfectly genuine. Fraud detection is the obvious case. I do not have a principled way to distinguish a legitimate gaming exemption from a convenient one, and I have not found anybody who does, which may mean the question is harder than the standard's drafters assumed or may just mean I have not looked in the right place.

References

SOURCES, WITH METHOD AND PROVENANCE WHERE IT BEARS ON THE WEIGHT A CLAIM CAN CARRY

Important claims were checked against regulators, audit bodies, public inquiries, standards organisations and primary official material. Where a source is research from a body with a commercial interest in the result, the reference says so. Several instruments described here were in passage or in draft when this was written, and dates should be re-checked before the material is relied on.

1. Information Commissioner's Office. *Guidance on AI and data protection*. Last updated 15 March 2023, carrying a notice that it is under review following the Data (Use and Access) Act; and *AI and biometrics strategy update*, March 2026, setting out the forthcoming automated decision-making guidance and statutory AI code of practice. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/
2. National Cyber Security Centre. *Guidelines for secure AI system development*. Version 1.0, 27 November 2023, published with CISA and 21 other international agencies. https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development
3. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1, 26 July 2024, with no subsequent revision; and *Concept Note: AI RMF Profile on Trustworthy AI in Critical Infrastructure*, page created 6 April 2026, last updated 17 July 2026, still a concept note. United States voluntary frameworks with no UK legal force. https://www.nist.gov/programs-projects/concept-note-ai-rmf-profile-trustworthy-ai-critical-infrastructure
4. Government Digital Service. *Artificial Intelligence Playbook for the UK Government*. Published 10 February 2025, last updated 9 September 2026. Recommends a use cases register for prioritising exploration and does not mandate an inventory of live AI systems. https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government/artificial-intelligence-playbook-for-the-uk-government-html
5. Government Digital Service. *Data and AI Ethics Framework*. Published 18 December 2025, successor to the Data Ethics Framework, published alongside a self-assessment tool. Guidance rather than a mandate. https://www.gov.uk/government/publications/data-ethics-framework/data-and-ai-ethics-framework
6. Government Digital Service. *Algorithmic Transparency Recording Standard: mandatory scope and exemptions policy*. Published 17 December 2024, last updated 9 September 2026; announced 6 February 2024; algorithmic transparency check added to digital spend controls February 2025. Record count of 152 taken from the published finder on 12 September 2026. A policy mandate with no statutory basis. https://www.gov.uk/algorithmic-transparency-records
7. Cabinet Office. *The UK Government Resilience Framework*. 19 December 2022, last updated 4 December 2023. https://www.gov.uk/government/publications/the-uk-government-resilience-framework
8. Cabinet Office. *UK Government Resilience Action Plan*, 8 July 2025; and *The UK Government Resilience Action Plan: 2026 Implementation Report*, 14 July 2026. https://www.gov.uk/government/publications/the-uk-government-resilience-action-plan-2026-implementation-report/the-uk-government-resilience-action-plan-2026-implementation-report-html
9. National Cyber Security Centre. *Operational Technology* collection, published and reviewed 18 March 2024; and *Secure connectivity principles for operational technology*, 14 January 2026, co-sealed by eight agencies across seven countries. https://www.ncsc.gov.uk/collection/operational-technology
10. National Cyber Security Centre. *Cyber Assessment Framework*, version 4.0, released 18 April 2024, reviewed 6 August 2025. Four objectives and 14 principles, none of which is AI-specific. https://www.ncsc.gov.uk/collection/cyber-assessment-framework
11. Department for Work and Pensions. *Artificial intelligence security policy*. Published 7 May 2025, last updated 7 August 2026. A departmental control imposed through the supply chain. https://www.gov.uk/government/publications/dwp-procurement-security-policies-and-standards/artificial-intelligence-security-policy
12. Department for Science, Innovation and Technology. *Responsible AI Toolkit*. Published 25 March 2024, last updated 15 November 2024. Almost every constituent item predates 2025; cite as an assurance-technique reference and not as current policy. https://www.gov.uk/government/collections/responsible-ai-toolkit
13. National Audit Office. *Use of artificial intelligence in government*, HC 612, 15 March 2024. Survey of 89 government bodies, autumn 2023, 98 per cent response rate from 87 bodies. Excluded simple rules-based automation and AI embedded in default tools. https://www.nao.org.uk/reports/use-of-artificial-intelligence-in-government/

14. Public Accounts Committee. *Use of AI in Government*. Eighteenth Report of Session 2024 to 2025, 26 March 2025. https://publications.parliament.uk/pa/cm5901/cmselect/cmpubacc/356/report.html

15. HM Treasury. *Treasury Minutes*, CP 1341, June 2025. Records the government commitment to publish records for in-scope tools identified as at March 2025 in departments, excluding arm's length bodies, by end 2025. https://assets.publishing.service.gov.uk/media/684fecc877c424182b0c4e5d/E03381954_CP_1341_Treasury_Minutes_Accessible.pdf

16. Parker, I., Studman, A. and Jones, E. *Learn fast and build things: lessons from six years of studying AI in the public sector*. Ada Lovelace Institute, 14 March 2025. An independent research institute. https://www.adalovelaceinstitute.org/policy-briefing/public-sector-ai/

17. Department for Science, Innovation and Technology. *State of digital government review*, CP 1251, January 2025. https://assets.publishing.service.gov.uk/media/678a47649752f24aa1573589/state-of-digital-government.pdf

18. Government Digital Service. *Microsoft 365 Copilot Experiment: cross-government findings report*. June 2025. 20,000 participants across 12 organisations, 30 September to 31 December 2024. No control group and no randomisation; headline time saving is self-reported and derived from the midpoint of categorical ranges. https://assets.publishing.service.gov.uk/media/683db42bd23a62e5d32680d0/M365_Copilot_Experiment_Findings_Report.pdf

19. HM Revenue and Customs. *Evaluating the impact of Microsoft Copilot in HMRC*, phase 3 trial evaluation report, 9 July 2026. Licences randomly allocated to 3,000 employees; outcome measure remains self-reported. https://www.gov.uk/government/publications/evaluation-report-phase-3-trial-of-microsoft-copilot/evaluating-the-impact-of-microsoft-copilot-in-hmrc

20. Makinson, L. *Evaluating Consult: an AI tool for enhanced public consultation analysis*. Incubator for Artificial Intelligence, Department for Science, Innovation and Technology, 14 May 2025. 5,354 question responses on a live Scottish Government consultation. The 62 per cent human benchmark was measured in internal testing on different data. No cost saving figure is given. https://ai.gov.uk/blogs/evaluating-consult-an-ai-tool-for-enhanced-public-consultation-analysis/

21. Hashem, Y., Bright, J., Chakraborty, S., Onslow, K., Francis, J., Poletav, A. and Esnaashari, S. *Mapping the Potential: generative AI and public sector work*. The Alan Turing Institute with the Office for National Statistics and the Department for Transport, June 2025. ONS Public Sector Time Use Survey, 1,513 respondents and 3,005 diary entries. A theoretical exposure estimate represented at its upper bound, not a measured productivity effect. https://www.turing.ac.uk/sites/default/files/2025-05/ons_tus_final_report.pdf

22. *Data (Use and Access) Act 2025*, section 80, substituting Articles 22A to 22D for Article 22 UK GDPR. In force 5 February 2026 by SI 2026/82. https://www.legislation.gov.uk/ukpga/2025/18/section/80

23. *R (Bridges) v Chief Constable of South Wales Police* [2020] EWCA Civ 1058, Court of Appeal (Civil Division), 11 August 2020. The appeal succeeded on three of five grounds. The public sector equality duty holding is at paragraph 199. https://caselaw.nationalarchives.gov.uk/ewca/civ/2020/1058

24. Joint Council for the Welfare of Immigrants and Foxglove, challenge to the Home Office visa streaming tool. Home Office conceded by letter 3 August 2020; use discontinued from 7 August 2020. Settled before any hearing, so there is no judgment and no precedent. https://www.foxglove.org.uk/2020/08/04/home-office-says-it-will-abandon-its-racist-visa-algorithm-after-we-sued-them/

25. Office for Statistics Regulation. *Ensuring statistical models command public confidence: learning lessons from the approach to developing models for awarding grades in the UK in 2020*. 2 March 2021. Examined all four UK qualification regulators. The report states no downgrade percentage. https://osr.statisticsauthority.gov.uk/publication/ensuring-statistical-models-command-public-confidence

26. Post Office Horizon IT Inquiry. *Volume 1*, HC 1119, presented 8 July 2025. Chair, Sir Wyn Williams. Figures at paragraphs 3.1, 3.8 to 3.12 and 3.24 to 3.25. Volume 1 covers human impact and redress; a progress update on the remainder was issued 8 July 2026. https://www.postofficehorizoninquiry.org.uk/reports-and-statements

27. District Court of The Hague. *NJCM and others v The Netherlands (SyRI)*, ECLI:NL:RBDHA:2020:865, judgment 5 February 2020. Held the SyRI legislation violated Article 8 ECHR and had no binding effect. https://www.rechtspraak.nl/organisatie-en-contact/organisatie/rechtbanken/rechtbank-den-haag/nieuws/2020/02/syri-legislation-in-breach-of-european-convention-on-human-rights

28. Berends, S. (Regioplan). *The Dutch childcare allowance affair*. European Commission ESPN Flash Report 2021/51, July 2021. Records 232 families in the initial 2017 report and more than 35,000 parents registered for compensation by April 2021. Wider figures in general circulation were not verified for this paper. https://ec.europa.eu/social/BlobServlet?docId=24723&langId=en

29. *The Network and Information Systems Regulations 2018*, SI 2018/506. Ten essential-service subsectors in Schedule 2; penalty bands at regulation 18. https://www.legislation.gov.uk/uksi/2018/506/contents

30. *Cyber Security and Resilience (Network and Information Systems) Bill*. Introduced 12 November 2025; all Commons stages complete; at report stage in the House of Lords as at early September 2026. Not an Act.

Commons Library briefing CBP-10442, 22 May 2026, last updated 8 July 2026. https://bills.parliament.uk/bills/4035

31. National Protective Security Authority. *Critical National Infrastructure*. Last updated 25 April 2023. The 13 CNI sectors and the official definition. https://www.npsa.gov.uk/critical-national-infrastructure-0

32. Cyber Monitoring Centre. *Statement on ransomware incidents in the retail sector*. 20 June 2025. Category two systemic event, estimated at £270m to £440m with no central estimate published; modelled from transaction-level data rather than audited loss. https://cybermonitoringcentre.com/2025/06/20/cyber-monitoring-centre-statement-on-ransomware-incidents-in-the-retail-sector-june-2025/

33. Information Commissioner's Office. *Fine of nearly £1m issued against South Staffordshire Plc and South Staffordshire Water Plc*. 11 May 2026. £963,900; 633,887 individuals affected. The notice concerns personal data and does not state that operational technology or supply was affected. https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2026/05/fine-of-nearly-1m-issued-against-south-staffordshire-plc-and-south-staffordshire-water-plc/

34. Drinking Water Inspectorate, disclosure under the Freedom of Information Act to Recorded Future News, reported November 2025. Five cyber attacks on UK water systems since 1 January 2024, not previously public because the network and information systems regime requires disclosure only where an attack causes disruption.

35. National Energy System Operator. *North Hyde Review: final report*. 2 July 2025. Records the elevated moisture reading in oil samples taken in July 2018 and 12 recommendations. https://www.neso.energy/national-energy-system-operator-publishes-final-report-review-north-hyde-substation-outage

36. Microsoft, statement of 20 July 2024 on the CrowdStrike content update, reporting 8.5 million Windows devices affected; and Kovrr, *The UK cost of the CrowdStrike incident*, 8 August 2024, modelling a UK cost of £1.7bn to £2.3bn. The Kovrr estimate is modelled rather than observed, assumes an average of one working day of downtime, discloses no organisational sample, and was produced and funded by a commercial cyber-risk quantification vendor. https://www.kovrr.com/reports/the-uk-cost-of-the-crowdstrike-incident

37. British Library. *Learning lessons from the cyber-attack: British Library cyber incident review*. 8 March 2024. Attack of 28 October 2023 attributed to Rhysida; approximately 600GB exfiltrated; 16 lessons. The review states that recovery costs were still being calculated and gives no figure. https://www.bl.uk/home/british-library-cyber-incident-review-8-march-2024.pdf/

38. Competition and Markets Authority. *Cloud services market investigation*. Opened 5 October 2023, final report and decision 31 July 2025; decision of 31 March 2026 not to prioritise strategic market status investigations into Microsoft's and Amazon's cloud infrastructure services, opening an investigation into Microsoft's business software ecosystem instead. https://www.gov.uk/cma-cases/cloud-services-market-investigation

39. Crown Commercial Service. *RM6200 Artificial Intelligence Dynamic Purchasing System*. A dynamic purchasing system rather than a closed framework. Supplier count and spend concentration were not obtainable from the published pages. https://www.crowncommercial.gov.uk/agreements/rm6200

40. Mutlu, U. and Kaewunruen, S. *Digitalised predictive maintenance in railways: a systematic review of AI, BIM and digital twins*. Infrastructures, volume 11, issue 3, article 87, 8 March 2026. PRISMA-compliant; 632 papers identified, 189 screened after deduplication, 73 included, 2015 to 2026. https://www.mdpi.com/2412-3811/11/3/87

41. Cabinet Office and Office for National Statistics. *Civil Service Statistics 2026*. Published 29 July 2026, reference date 31 March 2026. https://www.gov.uk/government/statistics/civil-service-statistics-2026/statistical-bulletin-civil-service-statistics-2026

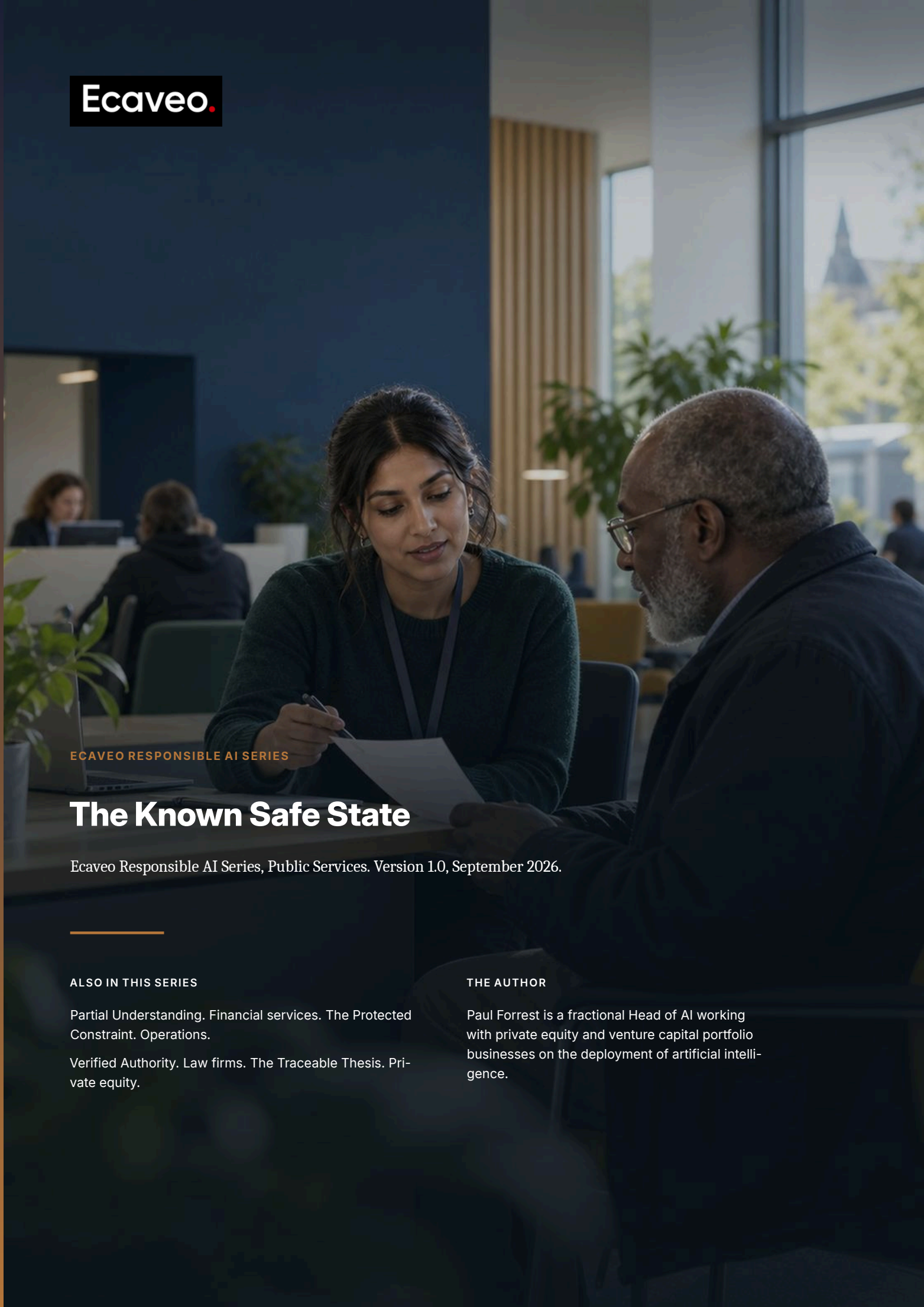
42. HM Revenue and Customs. *Annual Report and Accounts 2025 to 2026*. 9 July 2026. The £10bn attributed to artificial intelligence and advanced analytics is an undifferentiated departmental claim with no published methodology or counterfactual. https://www.gov.uk/government/publications/hmrc-annual-report-and-accounts-2025-to-2026/hmrcs-annual-report-and-accounts-2025-to-2026-executive-summary

43. National Audit Office. *DWP customer service*, HC 127, Session 2024 to 2025, 11 July 2024. https://www.nao.org.uk/wp-content/uploads/2024/07/dwp-customer-service.pdf

44. Ofwat. *Water Company Performance Report 2024-25*, 23 October 2025, and the leakage dataset of November 2025. Leakage of approximately 2,869 million litres a day across 17 companies, around 20 per cent of water put into supply, with 4 of 17 companies meeting their leakage targets. https://www.ofwat.gov.uk/regulated-companies/company-obligations/outcomes/water-company-performance-report-2024-25/

45. Parliamentary and Health Service Ombudsman. *Annual Report and Accounts 2024-25*, HC 998; and Local Government and Social Care Ombudsman. *Annual Report and Accounts 2024-25*, 18 December 2025. https://www.lgo.org.uk/information-centre/news/2025/dec/ombudsman-publishes-annual-report-and-accounts-2024-25

46. International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC 42001:2023, Artificial intelligence management system*. December 2023.



Ecaveo.

ECAVEO RESPONSIBLE AI SERIES

The Known Safe State

Ecaveo Responsible AI Series, Public Services. Version 1.0, September 2026.

ALSO IN THIS SERIES

Partial Understanding. Financial services. The Protected Constraint. Operations.

Verified Authority. Law firms. The Traceable Thesis. Private equity.

THE AUTHOR

Paul Forrest is a fractional Head of AI working with private equity and venture capital portfolio businesses on the deployment of artificial intelligence.