

From Pilot to Policy: Why Government AI Experiments So Rarely Scale

Word Count: 3536

Across governments, the gap between what artificial intelligence promises and what institutions have so far managed to do with it is widening. The UK Parliament's Public Accounts Committee has observed that government has no systematic mechanism for consolidating learning from AI pilots, and that successful at-scale adoption remains rare; a major Deloitte study across fourteen countries found that fewer than a third of organisations have scaled their Generative AI pilots or plan to.¹ This is far from the anticipated trajectory. Dunleavy and Margetts, in their account of the 'third wave' of Digital Era Governance, describe a future in which data science and AI (DSAI) enable "new options for partitioning state functions in ways that can maximise organisational productivity, in an 'intelligent centre, devolved delivery' model within vertical policy sectors," and make long-anticipated visions of personalised public services finally viable.² They are not naive about the timeline—they note that "the third wave changes set out here are likely to take two decades or more to be realized"—but the difficulty is no longer only one of pace.³ Pilots are not slowly maturing toward deployment; they are accumulating without progressing. The third wave that Dunleavy and Margetts describe is realistic given the technical possibilities but is not matched by the institutional realities that would carry it from experiment to routine operation.

Why, then, are so many government AI efforts unable to move beyond the pilot stage, or still stuck in germination? This essay will demonstrate that the binding constraint is not reducible

¹ OECD, *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions* (OECD Publishing, 2025), <https://doi.org/10.1787/795de142-en>, p. 83

² Patrick Dunleavy and Helen Margetts, "Data Science, Artificial Intelligence and the Third Wave of Digital Era Governance," *Public Policy and Administration* 40, no. 2 (2025): 185–214, <https://doi.org/10.1177/09520767231198737>, p. 185, 194.

³ Dunleavy and Margetts 2025, p. 206.

to technical feasibility or headline funding; it is organisational capacity, including the procurement, governance, and professional infrastructures through which funding and technology are converted into routine operation. Scaling GenAI requires not merely deploying tools, but transforming workflows, professional identities, and governance structures across whole institutions—changes that pilots rarely justify or compel. Pilots persist not only because they reduce risk, but because they can satisfy political demand for 'AI action' without triggering institution-wide disruption. While these dynamics apply to government AI use broadly, generative AI has intensified them; its lower technical barrier to entry has multiplied the volume of pilots without changing the institutional architecture that determines scaling potential. Three mechanisms in particular do most of the explanatory work, and this essay treats them in turn through cases drawn from the United Kingdom, France, and the United States: the insulation of pilot conditions, the structural discontinuity between pilot and routine operation, and the institutional rationality of indefinite non-progression.

The Pilot Scalability Paradox

A common tool in the governmental context for the introduction of generative AI has been pilots—bounded, time-limited experiments with stronger oversight, specific conditions, and a controlled environment that is deliberately insulated from institutional friction like procurement or grant requirements to test whether a certain technology or policy is effective. As the OECD notes, most governmental AI use cases "largely exist in the exploratory phase (e.g. proofs of concept, pilot projects) and have not yet been more broadly implemented or scaled beyond limited use."⁴ The problem has not been that these pilots have ended in failure. Rather, success is

⁴ OECD 2025, p. 83.

often a product of conditions that are artifacts of the pilot structure itself, with limited relevance to the frameworks that would govern them at scale.

Consider the three dimensions which Straub et al. identify as essential for legitimate AI use in government: operational fitness, which measures the degree to which the system in question matches the rules and standards of the organisation using it; epistemic alignment, which considers the degree to which public information about the AI application aligns with knowledge sharing requirements; and normative divergence, which considers the degree to which the software diverges from institutional standards and acceptable behaviour.⁵ The pilot setting creates a test environment with its own distinct dimensions, which, while similar to those of the original agency, are conducive to the piloting paradigm—the system is not necessarily subject to the same pressures across operational fitness, epistemic alignment, and normative divergence it would be expected to face in a scaled-up setting. As such, there consistently exists a misalignment between pilot and systemic pressures that are impossible to fully address.

These challenges are in part derivative of the very benefits provided by experimentation; as Selten and Klievink note, public organisations tend to navigate the tension between institutional requirements and innovation demands through separation.⁶ While this provides protection from friction such as traditionally burdensome procurement requirements, it also inadvertently protects against the sort of challenges that may reveal misalignment. That distance is structural: pilots are deliberately exempted from the procurement frameworks, data

⁵ Vincent J. Straub et al., “Artificial Intelligence in Government: Concepts, Standards, and a Unified Framework,” *Government Information Quarterly* 40, no. 4 (2023): 101881, <https://doi.org/10.1016/j.giq.2023.101881>, p. 8.

⁶ Friso Selten and Bram Klievink, “Organizing Public Sector AI Adoption: Navigating between Separation and Integration,” *Government Information Quarterly* 41, no. 1 (2024): 101885, <https://doi.org/10.1016/j.giq.2023.101885>, p. 1.

governance standards, legal accountability requirements, and professional role definitions that define routine operation—and those exemptions are constitutive of their success.

This dynamic is easily seen in the United Kingdom's NHS Artificial Intelligence Laboratory, the most ambitious systematic attempt to date to bridge the pilot-to-deployment gap in any health system. The Lab was launched jointly by NHS England and the Department of Health and Social Care in 2019 with an initial allocation of £250 million—reduced to £143.5 million in the 2022 spending review—and structured around an explicit phased architecture: Phase 1 to demonstrate technical and clinical feasibility, Phase 2 to develop prototypes and generate early efficacy data, Phase 3 to support real-world testing in care settings, and Phase 4 to support technologies with market authorisation but insufficient evidence for large-scale commissioning.⁷ Phase 5, by design, was the bridge, where projects would be expanded and rolled out at scale, and yet, as Cresswell et al. note, "No Phase 5 projects were funded."⁸ Eighty-six projects were funded across the preceding phases; the phase responsible for translating any of them into routine operation was vacated entirely.⁹

The failure of pilot progression was the result of a clear pattern: "Implementation and scaling were hindered by shifting objectives, limited capacity, and systemic misalignment with service needs," alongside decreasing funding as pilots failed to scale.¹⁰ The NHS AI Laboratory architecture had been designed precisely to confront the discontinuity between pilot conditions and routine deployment, and the discontinuity asserted itself regardless—not because the earlier phases failed, but because the insulation that made them tractable was inseparable from the

⁷ Cresswell, Kathrin, Robin Williams, Sheena Dungey, et al. "A Mixed Methods Formative Evaluation of the United Kingdom National Health Service Artificial Intelligence Lab." *Npj Digital Medicine* 8, no. 1 (2025): 448. <https://doi.org/10.1038/s41746-025-01805-w>, pp. 1-3.

⁸ Cresswell et al. 2025, p. 3.

⁹ Cresswell et al. 2025, p. 2.

¹⁰ Cresswell et al. 2025, p. 1.

fencing that made the scaling phase unaffordable when fiscal pressures arrived. The phases that could be insulated were funded; the phase that could not failed to gain further funding, in large part due to the insulation itself.

Structural Discontinuity and Designing for Scale

The greater challenge is that this insularity is multi-directional and reinforcing. Just as pilots allow for experimentation without the constraints of procedure, those same processes are insulated against the transformative changes required to make innovative use cases truly impactful. This is in large part the result of a substantive mismatch between agile government methodologies and traditional approaches in public sector project management. As Tai and Awasthi note, these mismatches are compounding across levels; while at the micro level conflicts are procedural and manageable, at the meso level (i.e., organisational management) they become structural, and at the macro level (i.e., governance) they harden into institutional blockages that require coordinated intervention to dislodge.¹¹ Effective scaling requires coordination across all three levels simultaneously, and it is precisely this coordination that pilot conditions cannot produce.

The same institutional surround that vacated Phase 5 reasserted itself in the AI Lab's component projects, even when the technology and the pilot itself were demonstrably successful. The clearest case is the AI Deployment Platform (AIDP), a pilot program for AI use within medical imaging services. The AIDP was specifically designed with an aim towards national rollout following a successful pilot, but when that rollout began to look unworkable, a senior

¹¹ Kuang-Ting Tai and Pallavi Awasthi, "An Exploration of Agile Government in the Public Sector: A Systematic Literature Review at Macro, Meso, and Micro Levels of Analysis," *Government Information Quarterly* 42, no. 4 (2025): 102082, <https://doi.org/10.1016/j.giq.2025.102082>, p. 13.

DHSC official reframed the non-progression in striking terms: that "a national rollout might not be the most appropriate route . . . a huge benefit because it saves what could have been . . . you know, a disaster."¹² The AIDP failure was not technical; the official was describing an institution that, having designed a pilot to test the case for scaling, used the pilot's run to construct a case against scaling. The deeper structural barrier was procurement itself. Cresswell et al. find that "many sites participating in the AI Awards had no transition plans to procure the implemented technologies after completion of the AI Awards, as they lacked evidence to support clinical or economic cases at the time that the AI Award funding was due to run out."¹³ As one supplier described it, "You just fall off the cliff edge at the end of it [the pilot] . . . you build a business case, and you've got the evidence, [but] there is no route to procure procurement."¹⁴ The capacity gap was reinforced by organisational reality: as one DHSC interviewee observed, "Staff that want to see improvements that have a real interest in research and AI simply do not have the capacity to engage" with the work that scaling would require, nor the funding and support to do so.¹⁵

This is what Giest and Klievink term innovation process creep, though here in inverted form: where their original concern was that organisational processes failed to register systems that had quietly become more intrusive than their formal architecture suggested, in the NHS AI Lab, the organisation could see its systems perfectly clearly, but the processes that would have absorbed them were never built. Their core diagnostic claim still holds in this inverted setting: as they put it, "ordinary organizational processes also need to co-develop with the innovation, for it

¹² Cresswell et al. 2025, p. 5.

¹³ Cresswell et al. 2025, p. 9.

¹⁴ Cresswell et al. 2025, p. 9. Quote includes bracketed insertions for clarity.

¹⁵ Cresswell et al. 2025, p. 10.

to be able to lead to performance gains."¹⁶ Procurement frameworks, governance pathways, and adopter-organisation workforce capacity remained calibrated to a pre-AI institutional environment, and the pilots completed within them were therefore not transmissible into routine operation, however technically successful they had been.

The dynamic Tai and Awasthi describe—procedural friction at the micro level hardening into structural blockage at the meso level and institutional blockage at the macro level—is visible in the AI Laboratory’s documentary record. At the micro level, there was a notable willingness to innovate: stakeholder communities, clinical champions, supplier partnerships, and formative evaluations, all with an eye towards scaling. At the meso level, that willingness collided with a 2022 spending review settlement that, in the AI Lab Board's own minutes, "required pausing and/or stopping a number of programmes."¹⁷ At the macro level, the procurement, accountability, and oversight architectures that would have given those programmes routes to operational continuation were never constructed. The compounding is precisely as Tai and Awasthi predict: pilots that the institution itself had designed as bridges instead became their own terminus.¹⁸

Hollowed Capacity and Recursive Bottlenecks

The rational institutional response to this discontinuity is to seek help bridging it—extending vendor relationships, retaining dedicated teams, and drawing on cross-sector collaboration. OpenAI and Google, among others, now promote major programs aimed at assisting governments and civil society partners with AI uptake. Yet external support cannot, by itself,

¹⁶ Giest & Klievink 2024, p. 391

¹⁷ Cresswell et al. 2025, p. 6.

¹⁸ Tai & Awasthi 2025, p. 13.

resolve a discontinuity that is structural rather than technical: the same institutional environment that defeats internal scaling attempts can defeat outsourced ones too.

This is not to deny that private–public partnerships can add value when structured to build internal capacity, but contracts are rarely structured to meaningfully do so, especially for black-box, highly valuable AI systems. As Mikhaylov et al. note, governments cannot deploy AI at scale without cross-sector collaboration, but this coordination carries structural features that make bridging discontinuity harder rather than easier.¹⁹ Knowledge asymmetries are fundamental: governments lack the internal expertise to evaluate what their partners are delivering, to interrogate model outputs, or to course-correct when systems drift. The same AI illiteracy that characterises internal deployment characterises the oversight of external deployment, complicating efforts to simply 'outsource' the challenge of scaling. The misaligned incentives of pilots mentioned in the first section only serve to compound this: since vendors are largely focused on maximising profits over public value, they are incentivised to demonstrate success within the pilot's protected conditions, not to engineer the institutional transformation that deployment necessitates. The result is that cross-sector collaboration produces paralysis at deployment scale: neither party has the tools to navigate institutional structures that were never designed for AI, and neither is incentivised to build them.

This is only further complicated by the fact that, as Wirtz et al. note, "government and public administration lag behind the rapid development of AI in their efforts to provide adequate governance."²⁰ This governance vacuum is not incidental, but structural. The procurement

¹⁹ Slava Jankin Mikhaylov et al., "Artificial Intelligence for the Public Sector: Opportunities and Challenges of Cross-Sector Collaboration," *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* 376, no. 2128 (2018): 1–21, p. 6.

²⁰ Bernd W. Wirtz et al., "The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration," *International Journal of Public Administration* 43, no. 9 (2020): 818–29, <https://doi.org/10.1080/01900692.2020.1749851>, p. 818.

frameworks, legal requirements, and oversight mechanisms that would need to govern deployed AI were designed for a pre-AI institutional environment. Pilots can operate in this vacuum because they are exempted from it; deployment cannot, because it isn't. Unfortunately, the lack of governance infrastructure means that there are no load-bearing administrative structures for advanced AI adoption, further obfuscating the potential for uptake.

This is all to say that the dynamic capabilities that would bridge the discontinuity between pilot and comprehensive deployment are properties of the institution itself, not simply procurable services. As Kattel notes, "outsourcing key functions to external companies [only] hollowed out public agencies of core digital skills."²¹ The wall is not between the institution and the vendor; it is between the sanitised world of the pilot environment and the institutional realities and resistance to comprehensive transformation necessary for realistic deployment, whether internal or external, of any new AI system.

The recursive nature of this hollowing is visible in the United States Internal Revenue Service, where a generative AI tool designed to ease procurement bottlenecks has itself become trapped in the procurement bottleneck it was built to relieve. The AI Contract Document Toolbox (AICDT), described in the U.S. Internal Revenue Service itself as a generative AI tool "designed to assist with procurement document tasks, streamlining the process and improving overall efficiency,"²² is a textbook instance of the discontinuity Kattel describes. Its operational role is straightforward: IRS employees upload draft solicitation documents, contractor proposals, and federal acquisition policy questions, and the tool drafts evaluations, redlines, and policy responses for human review. The technology works, and yet, as of October 2024, the agency

²¹ Rainer Kattel, *Dynamic capabilities of the public sector: Towards a new synthesis* (Instituto de Pesquisa Econômica Aplicada, 2022), <https://doi.org/10.38116/dp267>, pp. 22-23.

²² Internal Revenue Service, *Privacy and Civil Liberties Impact Assessment: AI Contract Document Toolbox (AICDT)*, October 2024, p. 1.

classified the system as "Initial proof of concept and experimentation with different AI solutions. Which may later grow into a pilot project."²³ Even after extensive testing, the project remains in its infancy.

This pre-pilot status is not anomalous within the broader IRS portfolio: the Government Accountability Office's March 2026 audit found that the agency's AI inventory had grown from ten use cases in 2022 to 126 by June 2025, of which 77 remained in development.²⁴ The proximate cause of the stagnation, the GAO documents, is the same dynamic Kattel identifies. The IRS's Office of the Chief Procurement Officer "lost over 40 percent of its staff between May and June 2025," and "about 80 percent of IRS's use cases in June 2025 involved some degree of contracting."²⁵ Procurement officials warned the GAO that the resulting workforce gap "could result in significant delays in entering into new contracts and modifying or renewing existing ones, including contracts that support mission-critical programs,"²⁶ and the GAO's overarching conclusion was unambiguous: "Without staff to support AI capabilities and oversee individual use cases, IRS is at risk of initiating projects the agency does not have resources to support or deploying AI systems without resources to ensure they are used responsibly."²⁷ The recursion is exact: AICDT was built to scale procurement, and procurement is precisely the meso-level capacity that has been hollowed out beneath it, resulting in an ironic circumstance in which the very tool to help resolve the problem becomes stuck within it.

²³ Internal Revenue Service 2024, p. 3.

²⁴ US Government Accountability Office, *Artificial Intelligence: IRS Actions Needed to Address Skills Gaps, Information Quality, and Strategic Management*, GAO-26-107522, March 2026, p. 12.

²⁵ US Government Accountability Office 2026, p. 25.

²⁶ US Government Accountability Office 2026, p. 25.

²⁷ US Government Accountability Office 2026, p. 29.

Deferred Reckoning

These challenges are not the product of AI pilots themselves so much as the latent problems beneath them; the pilot process defers those problems through controlled conditions, while structural discontinuity leaves limited internal capacity to respond and prior outsourcing hollows out the skills needed to adapt. Stagnation, then, is not some new failure, but the point at which accumulated weaknesses become visible.

The reason these issues accumulate rather than resolve is not incapacity alone, but a structural unwillingness to create the conditions under which resolution would be possible. As Madan and Ashok note, experimentation is institutionally rational: it allows organisations to signal responsiveness to AI demand without committing to the organisational transformation that deployment would require.²⁸ In this sense, pilots do not merely defer the challenges identified above—they serve a legitimisation function, producing sufficient visible activity to satisfy political demand while leaving underlying architecture untouched. The OECD’s finding that most AI use cases remain at proof-of-concept stage is therefore less evidence of a slow pipeline than of one that often terminates there by design. This is reinforced by the governance vacuum Wirtz et al. identify: procurement, legal accountability, and oversight mechanisms were built for a pre-AI environment, leaving few formal pathways through which the case for scaling can even be made.²⁹

The clearest illustration of this institutional rationality is Albert, a generative AI assistant developed in-house by the French Direction interministérielle du numérique (DINUM) and

²⁸ Rohit Madan and Mona Ashok, “AI Adoption and Diffusion in Public Administration: A Systematic Literature Review and Future Research Agenda,” *Government Information Quarterly* 40, no. 1 (2023): 101774, <https://doi.org/10.1016/j.giq.2022.101774>, p. 12.

²⁹ Wirtz et al. 2020, p. 818.

trialled at forty-eight France Services centres from 2024.³⁰ The administration had committed publicly to scaling, anticipating that the system would scale over the course of 2025 and that its use would be generalised across other State services. In January 2026, DINUM announced that Albert would *not* be scaled in its current form.³¹ Rather than treating the failure as a verdict on Albert, DINUM reframed the decision as one of continuity, assuring the public that Albert had not been abandoned. The agency described the underlying project as having evolved into a more powerful successor, called Assistant IA, integrating Mistral AI models, and expected to be piloted with ten thousand public agents across government ministries until June 2026.³² The institutional response to a pilot that did not scale was, in other words, a larger pilot.

Most striking is DINUM's explicit framing of what the new pilot is for: it must, according to the agency, establish the specific cost of generalisation across ministries.³³ The deferral is recursive in the institution's own voice: a pilot whose justification for existing is to assess whether its predecessor's scaling would have been affordable. Madan and Ashok's claim that experimentation is institutionally rational—that it allows organisations to signal responsiveness to AI demand without committing to the transformation that deployment would require—here finds expression in DINUM's own administrative language. France's case demonstrates that the dynamic does not require the workforce hollowing of the IRS or the funding-cycle vacation of the AI Lab; it requires only an institution capable of reclassifying non-progression as continued progress, and the bureaucratic willingness to commission a successor pilot whose terms of reference encode the deferral itself.

³⁰ Agence France-Presse / *Weka*, "Albert, l'outil d'IA générative, expérimenté à France Services ne sera pas généralisé," *Weka*, 12 January 2026, <https://www.weka.fr/actualite/administration/article/albert-l-outil-d-ia-generative-experimente-a-france-services-ne-sera-pas-generalise-209194/>.

³¹ AFP / *Weka* 2026. AI was used to assist with translation from French to English.

³² AFP / *Weka* 2026. AI was used to assist with translation from French to English.

³³ AFP / *Weka* 2026. AI was used to assist with translation from French to English.

The result is a self-reinforcing equilibrium: pilots produce incremental, bounded value that does not threaten existing structures; that bounded value does not justify the disruption of scaling; and the absence of scaling means institutions never develop the absorptive capacity that scaling would require. Unless the cycle is broken by deliberate political and organisational willingness, Dunleavy and Margetts' third-wave vision cannot materialise—not for technical reasons, but because the institution has chosen to remain structurally unchanged. Stagnation, then, is not failure; it is the system working precisely as the institutional environment has forced it to.³⁴

Conclusion

This essay asked why so many government GenAI efforts are behind schedule or fail to scale beyond pilots, and argued that the binding constraint is not reducible to technical feasibility or headline funding; it is organisational capacity, including the procurement, governance, and professional infrastructures through which funding and technology are converted into routine operation. The first section demonstrated how ‘success’ in insulated, exempted conditions is misread as evidence of deployability. The second section considered how, when projects move toward routine operation, structural discontinuity reasserts itself across micro, meso, and macro levels as agile experimentation collides with procurement, accountability, data governance, and professional role architectures built for a pre-AI state. As the third section illustrated, private–public partnerships may rationally appear to bridge the gap, but they cannot resolve a discontinuity that is structural rather than technical and can further hollow out internal capacity. In this sense, pilots do not solve these problems so much as defer them—until a point at which

³⁴ Kattel 2022, p. 23; Dunleavy and Margetts 2025, p. 206.

the accumulated institutional weaknesses become visible. The three cases examined here illustrate this deferral across distinct mechanisms: the NHS AI Lab's vacated scaling phase, where a programme architecture explicitly designed to bridge the discontinuity simply omitted the bridging step; the IRS's recursive procurement bottleneck, where a generative AI tool built to ease procurement constraints sits stuck behind exactly those constraints; and DINUM's response to Albert's non-progression, where the institutional answer to a pilot that did not generalise was a larger pilot whose evaluation must establish the cost of generalisation. The mechanism varies—a vacated phase, a hollowed workforce, a reclassification—but the equilibrium does not.

This is not to imply that pilots are pointless or that Dunleavy and Margetts' third-wave vision is unattainable; pilots built to confront routine governance constraints—procurement, accountability, professional roles—instead of bypassing them could, in principle, generate the signals that justify scaling, and they remain a useful measure of feasibility from the technical perspective. Rather, it is to suggest that true, transformative advancement will require a willingness to transition beyond the analogue norm, and whether state bureaucracies can make that transition remains to be seen.

Works Cited

- Agence France-Presse. “Albert, l’outil d’IA générative, expérimenté à France Services ne sera pas généralisé.” *Weka*, January 12, 2026.
<https://www.weka.fr/actualite/administration/article/albert-l-outil-d-ia-generative-experimente-a-france-services-ne-sera-pas-generalise-209194/>.
- Cresswell, Kathrin, Robin Williams, Sheena Dungey, et al. “A Mixed Methods Formative Evaluation of the United Kingdom National Health Service Artificial Intelligence Lab.” *Npj Digital Medicine* 8, no. 1 (2025): 448. <https://doi.org/10.1038/s41746-025-01805-w>.
- Dunleavy, Patrick, and Helen Margetts. “Data Science, Artificial Intelligence and the Third Wave of Digital Era Governance.” *Public Policy and Administration* 40, no. 2 (2025): 185–214. <https://doi.org/10.1177/09520767231198737>.
- Giest, Sarah N., and Bram Klievink. “More than a Digital System: How AI Is Changing the Role of Bureaucrats in Different Organizational Contexts.” *Public Management Review* 26, no. 2 (2024): 379–98. <https://doi.org/10.1080/14719037.2022.2095001>.
- Internal Revenue Service. *Privacy and Civil Liberties Impact Assessment: AI Contract Document Toolbox (AICDT)*. U.S. Department of the Treasury, Internal Revenue Service, 2024.
- Kattel, Rainer. *Discussion Paper 267*. Instituto de Pesquisa Econômica Aplicada, 2022.
<https://doi.org/10.38116/dp267>.
- Madan, Rohit, and Mona Ashok. “AI Adoption and Diffusion in Public Administration: A Systematic Literature Review and Future Research Agenda.” *Government Information Quarterly* 40, no. 1 (2023): 101774. <https://doi.org/10.1016/j.giq.2022.101774>.
- Mikhaylov, Slava Jankin, Marc Esteve, and Averill Champion. “Artificial Intelligence for the Public Sector: Opportunities and Challenges of Cross-Sector Collaboration.” *Philosophical*

Transactions: Mathematical, Physical and Engineering Sciences 376, no. 2128 (2018): 1–21.

OECD. *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions*. OECD Publishing, 2025. <https://doi.org/10.1787/795de142-en>.

Selten, Friso, and Bram Klievink. “Organizing Public Sector AI Adoption: Navigating between Separation and Integration.” *Government Information Quarterly* 41, no. 1 (2024): 101885. <https://doi.org/10.1016/j.giq.2023.101885>.

Straub, Vincent J., Deborah Morgan, Jonathan Bright, and Helen Margetts. “Artificial Intelligence in Government: Concepts, Standards, and a Unified Framework.” *Government Information Quarterly* 40, no. 4 (2023): 101881. <https://doi.org/10.1016/j.giq.2023.101881>.

Tai, Kuang-Ting, and Pallavi Awasthi. “An Exploration of Agile Government in the Public Sector: A Systematic Literature Review at Macro, Meso, and Micro Levels of Analysis.” *Government Information Quarterly* 42, no. 4 (2025): 102082. <https://doi.org/10.1016/j.giq.2025.102082>.

U.S. Government Accountability Office. *Artificial Intelligence: IRS Actions Needed to Address Skills Gaps, Information Quality, and Strategic Management*. U.S. Government Accountability Office, 2026.

Wirtz, Bernd W., Jan C. Weyerer, and Benjamin J. Sturm. “The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration.” *International Journal of Public Administration* 43, no. 9 (2020): 818–29. <https://doi.org/10.1080/01900692.2020.1749851>.

