Adoption before proof

Government AI is back-office text work, and its adoption figure counts one touch in 30 days

Government AI today is internal text work at low volume, and the adoption rate published beside it counts one interaction in 30 days.

Most of what government runs as AI never reaches a citizen. Of 421 classifiable Dutch central-government systems, 281 have no direct effect on citizens or businesses, and the largest single class, 124 systems, is knowledge processing: search over internal documents, transcription and anonymisation[1]. Volume is thin. Only 120 of the 433 systems reported are in use, and 88 per cent of the 70 bodies surveyed run three systems or fewer[1]. Of the 32 UK bodies with deployed AI, six do anything that faces the public[2].

The adoption rate governments publish counts a licence holder who touched the tool once in 30 days. On that definition, 20,000 licences across 12 UK organisations peaked at 83 per cent[3]. The department holding 1,000 of those licences published telemetry for the same tool and the same quarter, and recorded 64 per cent weekly use, 21 per cent daily use, and 1.14 tool actions per user per working day[4].

A business case takes its population from recurring use and its rate from a measured saving. Re-running the best documented case at the weekly rate cuts the multiplier by about a fifth, and by three quarters if daily use is the right base[3,4]. Six steps separate a licence count from cash: licences issued, 30-day reach, recurring use, self-reported time, output scored against non-users, and a budget line that falls. No administration has reached the last of those for an AI system, which needs a baseline, its measurement date, and the establishment or vote the money leaves.

Three questions dispose of any adoption claim: the window, the denominator, and which of those six steps the number sits on. A figure with no stated window is 30-day reach, and multiplying it by a per-user time saving produces a budget number the evidence does not carry. The one published costed projection does exactly that: about £50m a year from up to 50,000 licences and a self-reported saving[5].

RECURRING USE

64%

weekly users in the department holding 1,000 of 20,000 licences, against an 83 per cent adoption headline on the same tool and quarter[4]

PUBLIC-FACING DEPLOYMENTS

6 of 32

UK bodies with deployed AI that do anything facing the public; 21 use it to support operational decision-making[2]

SYSTEMS PER BODY

88%

of 70 Dutch central-government bodies run three systems or fewer, and 120 of 433 reported systems are in use[1]

HOW WE KNOW

The 30-day window is the vendor telemetry default and is stated as such in the UK cross-government report, so a published adoption rate counts licence holders who logged one interaction in a month[3]. Weekly and daily series exist for one department, produced from telemetry it already held[4]. Every figure here covers a generic office assistant. No recurring-use figure has been published anywhere for a case-processing system.

The legal basis stops as many projects as the model does

Both causes are answerable before any code is written: whether a lawful basis exists for running the system, and whose budget pays in year two.

Failure arrives before production. Of 40 publicly funded Danish AI projects, six reached operation and 16 finished without ever operating; four of those 16 stopped on the legal basis and four on performance, and legal difficulty was reported by 68 per cent of the participating municipalities and regions[6]. Several projects had never put operating and retraining costs in the business case, and several with good results stopped because the receiving organisation ranked other candidates above them for its own money[6].

Conversion does not order the adoption waves. Across 2,133 US federal use cases, the only inventory large enough to test the rule, internal support converted to operation at 33.6 per cent, government services at 45.1 per cent, and the declared rights-impacting class at 74.2 per cent[7]. What internal work buys is elapsed time, a median of 6.4 months to production against 12.1 months for citizen-facing cases[7]. Read that with its defect: the same internal class appears at 25.1 and at 45.2 per cent depending on which agency did the labelling[7].

Name the postponed volume in the same paper as the sequence. One administration has priced both halves, a document-conversion tool given free to every English planning authority at an estimated 250,000 officer-hours a year, against £8.2m for a case-triage service covering three councils[8]. The boundary is the record system. Document work runs on the estate as it stands, and the UK auditor holds that building AI on top of existing systems without fixing them limits what government can reach[2].

Four signatures before any code: the lawful basis, the service owner, the case system, the year-two line.

Four of the sixteen Danish projects that never operated stopped on the legal basis[6].

REACHED OPERATION

6 of 40

publicly funded Danish AI projects; 16 finished without ever operating, four of them stopped on the legal basis[6]

TIME TO PRODUCTION

6.4 months

median for internal-support use cases against 12.1 months for citizen-facing ones, across 2,133 US federal cases[7]

CASE-TRIAGE CONTRACT

£8.2m

for a three-council alpha, against a document-conversion tool released free to every English planning authority[8]

No evaluation has both a comparison group and a measure other than self-report

Government has no measured estimate of what office AI saves: every published benefit figure is staff estimating their own time, including the one that has a comparison group[9].

The cross-government figure of 26 minutes a day comes from 7,115 banded self-reports, 36 per cent of licence holders, converted at band midpoints, with 17 per cent of respondents reporting no clear saving at all. The same report states that experimental constraints made it impossible to identify how the time saved was spent[3].

Each component of the missing trial has already run, in a different department. One randomised 3,000 of its 3,500 licences and then surveyed only licence holders, reporting about 60 minutes a week after a 20 per cent cut for non-use[5]. Another built the quantitative untreated arm, 1,716 users against 2,535 non-users drawn by stratified random sample, and returned 19 minutes a day with a 95 per cent interval of 17 to 22[9]. The 19 minutes is not a corrected version of the 26: a comparison group fixes the counterfactual and leaves self-report bias untouched.

Scoring produced output against matched non-users reversed the diary on two of four tasks. Data analysis, claimed in the diary at 0.6 hours saved, took observed users 25:01 against 20:33 and scored 1.5 out of 5 on accuracy and quality against 2.7. Slide production, claimed at no saving at all, came 7 minutes 43 seconds faster while scoring 1.5 on accuracy against 5. Only report summarising held on both measures, at 12:37 against 41:34[4].

A flat percentage discount therefore corrects nothing, because the bias changes sign by task. Score accuracy and quality alongside time on any observed exercise, and name the task mix any diary covers: a business case weighted towards analysis is claiming a saving the one direct measurement found negative.

WITH A COMPARISON GROUP

19 min

a day, 95 per cent interval 17 to 22, against a 26-minute cross-government headline on the same tool, from a different population and a different survey[3,9]

RANDOMISED, THEN UNMEASURED

3,000 of 3,500

licences allocated at random, after which only licence holders were surveyed, so the allocation bought balance and no comparison[5]

OBSERVED DATA ANALYSIS

25:01

taken by users against 20:33 by matched non-users, scoring 1.5 out of 5 on accuracy against 2.7, on a diary claim of 0.6 hours saved[4]

HOW WE KNOW

The observed-task cells are two and three people per arm across 11 sessions, blind-scored on a five-point scheme by the researcher who did not run the session[4]. The tasks were fictional exercises completed within an hour, so they test capability under observation. A department ran a comparable exercise with 49 participants and published no results from it[5].

Seventeen of seventeen departments booked workforce efficiencies with no published derivation

The first date any of these figures can be tested is Spending Review 2027, two years after they entered the plans[10].

All 17 main UK departments specified workforce-reform efficiencies for 2026-29 in the plans published with Spending Review 2025, most attributing them in part to digital and AI tools. The auditor reported in July 2026 that the plans give no details of how departments derived the figures, and that the efficiency guidance does not say how to apply its general rules to workforce efficiencies[10].

The calculation rule already exists. A cash-releasing saving must rest on baseline cost information, a counterfactual spending profile and outturn spending data; it must be realised inside the financial year, release cash from the department's budget net of costs, and be readily defendable to the auditor[11]. What is missing is an obligation to show the working before the number enters a settlement.

The one published reconciliation of an automation saving found 9 of the 10 highest-claiming bots had misstated their hours. The audit put three questions to end users, how often they perform the task, how long it takes and whether the tool saves any time, across a sample covering 107,242 of more than 240,000 claimed hours. One bot credited with 3,650 hours saved users no time at all, two carrying 15,000 hours were retired after four months and booked for a full year, and 74 per cent of bot owners reported no cost avoidance[12]. The estimating method behind those errors, headcount times frequency times an assumed duration, is the method in use for AI today.

Time released is not cash until a line falls. Over 70 per cent of users in the cross-government experiment moved time out of search and mundane tasks into other work, and the evaluation could not say where it went[3]. Name the establishment or budget line at business-case stage; where no line can be named, record the saving as non-cash-releasing and stop multiplying it by a paybill rate.

DEPARTMENTS BOOKING WORKFORCE EFFICIENCIES

17 of 17

main UK departments specified 2026-29 workforce efficiencies, most attributed partly to digital and AI tools, none with a published derivation[10]

AUTOMATION CLAIMS MISSTATED

9 of 10

highest-claiming bots differed from their estimate when users were re-asked three questions; one credited with 3,650 hours saved no time[12]

FIRST TESTABLE DATE

SR 2027

Spending Review 2027, when departmental efficiencies booked for 2026-29 are next revisited[10]

WHAT CASH-RELEASING REQUIRES

On the finance ministry's own framework a cash-releasing saving needs baseline cost information, a counterfactual spending profile and outturn data; it must be realised within the financial year and release cash from the department's budget net of costs, at equal or greater value in every later year[11]. Reinvested time meets none of those tests. A department that cannot name the establishment line its efficiency came out of has, on that framework's terms, not reported an achieved saving.

A challenge count is not an error monitor

Defective automated decisions do not surface as appeals, so an appeal rate is not evidence that a system is working.

The one administration that reviewed the decisions nobody had appealed found nine defects for every one the challenge channel surfaced. A review of 62,784 unemployment fraud penalties reversed 85 per cent of the 40,195 findings the computer had resolved and 44 per cent of the 22,589 an investigator had handled, roughly 44,000 defective determinations and $20.8m refunded, while appeals over the same window ran at 4,955 of 67,739 cases, 7 per cent, the same rate as human-decided disability benefit decisions[13]. The computer-resolved pile was wrong about twice as often as the investigator-handled one, and one appeal rate covers both.

Per-decision entitlements exist with no duty to count them. The Canadian directive, in force since 1 April 2019, requires plain-language notice at every impact level, the principal factors behind a denial from Level II, a human final decision at the top two levels, and recourse that is timely, effective and easy to access, and attaches no counting duty to any of it[14]. A decision register needs three counters, published per system per year: decisions issued, challenges lodged, outcomes changed.

The redress channel is mostly fixed cost, so both directions of a business case are wrong. Cost per disposal in one social security tribunal ran at £1,091 in 2021-22 against £257 in 2013-14, disposals having fallen 83 per cent while total cost fell 28 per cent[15]. A case pricing new challenges at the current average overstates what they cost to absorb, and a case claiming savings from suppressing challenges overstates those by the same arithmetic.

Two lines belong in the first year's budget. Fund a random audit of decisions nobody challenged, sized from a sampled defect rate and reporting outside the programme, because that is the only method that found the defects above. Ask the tribunal for the marginal cost of an additional hearing before any rollout that changes decision volumes, and size the review channel from decision volume rather than a borrowed appeal rate.

DEFECTS THE APPEAL RATE MISSED

44,000

unappealed determinations reversed on review, with $20.8m refunded, against 4,955 appeals lodged over the same window[13]

APPEAL RATE

7%

4,955 of 67,739 cases over the window in which those defects arose, the same rate as comparable human-decided benefit decisions[13]

COST PER TRIBUNAL DISPOSAL

£1,091

in 2021-22 against £257 in 2013-14, disposals having fallen 83 per cent while total cost fell 28 per cent[15]

HOW WE KNOW

Both benchmarks come from rule-based automation and from human-decided benefit appeals[13,15]. They bound a plausible contest rate for an AI-assisted service and predict none. The tribunal figure is an estimated average across all case types, and the minister answering warned it is not a true indication of unit cost by type.

3.1

How should the government establish whether its office AI is working?

No published evaluation has both a comparison group and a measure other than self-report. The cross-government figure of 26 minutes a day rests on banded self-report from licence holders[3,9]. Three departments have each run one component of the missing trial, and none has combined two[4,5,9].

	A · Withheld-licence quarter	B · Never-licensed survey	C · Scored observed tasks
BENEFITS	A comparison group and a measure beyond self-report, at no licence cost.	A comparison group at survey cost; 19 minutes a day, 17 to 22[9].	About two analyst-weeks, and the only component that measures quality[4].
TRADE-OFFS	Delays staff who see colleagues already holding the tool.	Selection survives the controls, and both arms self-report.	Produces no departmental figure.
RISKS	Legally untested; no published ruling permits or bars it.	Reads as a debiased headline and is not one.	Cells of two and three people get quoted as the finding[4].
CONDITIONS	Staff consultation completed before allocation, on pain of €1,000 a day[16].	Publish the response rates, 48 per cent for users and 27 per cent for non-users[9].	Publish cell sizes beside every figure.

RECOMMENDATION

Adopt Option A. The permanent secretary commissions one quarter combining random allocation of the next licence tranche, a survey of the never-licensed remainder and blind-scored observed tasks, publishing accuracy and quality scores beside the times.

OWNER · PERMANENT SECRETARY, DEPLOYING DEPARTMENT

DECIDE BY · Q1 2027, BEFORE COVERAGE PASSES A THIRD OF HEADCOUNT

RESIDUAL RISK · FOREGONE BENEFIT

Once coverage nears headcount a withheld seat costs roughly £250 a quarter. That per-seat value divides a £50m projection across up to 50,000 licences, and rests on the self-reported figure the trial exists to test[5].

3.2

Out of whose budget does year two come, and when is that settled?

Whole-of-government AI licences were bought at $1 an agency for a year, and no vendor has published a renewal price[17]. The agreements expire on 30 September 2026, so the second-year cost is a procurement event[18].

	A · Repayable central fund	B · Promotional deal, price fixed at award	C · Departmental base from day one
BENEFITS	Forces the year-two question at application, with repayment from the receiving body's own base.	Buys adoption at scale for almost nothing, and builds twelve months of usage to price against[17].	No cliff, because there was never a subsidy.
TRADE-OFFS	Repayment matures years before the savings: $13.5m realised against $1.03bn invested, 1.3 per cent[19].	Renewal falls at maximum switching cost, twelve months after signature[18].	Slower, uneven adoption, with the largest caseloads served last.
RISKS	98.3 per cent of the portfolio's expected savings fall due in FY2027 or later[19].	No forward plan can carry a year-two number that does not exist.	No central purchase means no central telemetry, so the usage series is assembled body by body[3].
CONDITIONS	Only where the base can absorb the run-rate.	List price and costed exit in writing at award.	Licence prices a department can absorb.

RECOMMENDATION

Adopt Option B with both controls. The central procuring body takes the vendor's post-promotion per-user list price in writing at award and prices a costed exit on the same date. It refuses central money for any system whose year-two budget line is not named.

OWNER · CENTRAL PROCURING BODY, WITH THE FINANCE MINISTRY

DECIDE BY · BEFORE THE CURRENT AGREEMENTS EXPIRE, 30 SEPTEMBER 2026

RESIDUAL RISK · RUN-RATE

No state publishes what an operational government AI system costs in its second year. Danish projects with good results stopped because operating and retraining money had never entered the business case[6].

3.3

How is a person told of an AI-assisted decision, and given a human to look again?

GDPR Article 22(2)(b) leaves the safeguard for a decision authorised by statute to the statute itself[20]. The AI Act gives an explanation right from 2 August 2026, sixteen months before the deployer's duty to tell anyone a system was used, moved to 2 December 2027 for the high-risk uses the Act lists, which include deciding eligibility for public benefits[21,22].

	A · Amend the enabling statute	B · One horizontal amendment	C · Notice and review as policy
BENEFITS	The only route that guarantees review for a statutory decision.	One drafting exercise, and it reaches systems already deployed.	Available immediately, on a graded model in force since 2019[14].
TRADE-OFFS	Slow and per-Act: a dozen Acts means a dozen amendments.	A generic clause may not discharge the specific duty.	Policy is revocable and confers no enforceable right.
RISKS	The amendment lands after the system does.	An estate-wide entitlement, challenge volume unestimated.	A lapsed notice policy leaves the explanation right empty.
CONDITIONS	Pair the safeguard with the counting duty in the same clause.	Only where the general statute binds every deploying body.	Treat as a bridge to a statutory amendment with a named date.

RECOMMENDATION

Adopt Option A, with Option C running from go-live as the bridge. The sponsoring ministry amends the enabling statute to carry human review and notification, and writes into the same clause a duty to publish decisions issued, challenges lodged and outcomes changed per system per year.

OWNER · SPONSORING MINISTRY

DECIDE BY · BEFORE GO-LIVE; FOR SYSTEMS ALREADY DEPLOYED, BEFORE 2 DECEMBER 2027

RESIDUAL RISK · WAITING

Public-authority deployers of high-risk systems already placed on the market have until 2 August 2030[21,22]. The systems with the longest operational history carry the longest statutory run.

Four numbers no administration publishes

Four figures are missing everywhere: what an operational system costs in year two, what the same staff would have produced without the tool, how many of its decisions are contested, and what one assurance post produces in a year.

No published business case separates record and data preparation from integration into a named case system from model build with a five-year run cost. Officials on one US Army programme told auditors that buyers price model training and overlook the infrastructure needed to sustain the system, so agencies may acquire what they cannot afford to run; across the 13 acquisitions reviewed, officials at all five bodies reported AI costs hard to understand[23]; US budget guidance has required operating and maintenance costs in AI solicitations since April 2025[24]. The nearest prices are contract totals, £8.2m for a three-council case-triage service among them, broken down nowhere[8]. Test a first bid against that figure, and refuse any case that arrives as a single line.

No evaluation has combined an untreated comparison group with a measure other than self-report, and no department has held licences back and scored produced output in the same quarter[4,5,9]. Until one does, every benefit figure in circulation is staff estimating their own saved time, and the gap between 26 minutes a day and about 60 minutes a week on the same tool stays undecomposed[3,5].

No rule requires a body to publish decisions issued, challenges lodged and outcomes changed, and no register carries the three figures for any system[14]. A contest rate borrowed from a neighbouring body carries no information about the new one, so collect the counts from the first year of operation and report them outside the programme.

Oversight is budgeted in posts, and no administration publishes what a post produces in a year. The European Commission priced third-party review of one high-risk system's documentation at one to two and a half days, €3,000 to €7,500, so a 220-day assessor year buys 88 to 220 files at most[25]. Thirty per cent of 87 UK bodies had assurance processes explicitly covering AI risks, and about a third of digital and data professionals were contractors, agency or temporary staff[2]. Write the oversight line as expected volume times unit price with the utilisation assumption stated.

LIVE SYSTEMS UNASSESSED

35%

42 of 120 Dutch central-government AI systems in use whose owner did not know whether it was meeting expectations, rising to 72 per cent for ongoing experiments[1]

ASSURANCE COVERING AI

30%

of 87 UK bodies had risk and quality assurance processes explicitly incorporating AI risks; 15 of the 32 with AI deployed held no register of their own live use cases[2]

HOW WE KNOW

Both figures are self-reported surveys: 87 UK government bodies in autumn 2023, and all 70 Dutch central-government organisations in 2024. Neither auditor independently checked what bodies submitted, and the UK figures predate the 2025 adoption push, so 30 per cent is a floor. No equivalent survey series exists for a small state, so these two administrations are the whole evidence base.

Seven instructions, each with an owner and a deadline

The first deadline is 30 September 2026.

The finance ministry amends the efficiency guidance: no AI-attributed efficiency enters a settlement without a baseline, its measurement date, and the budget or establishment line the saving leaves[10,11]. Where none can be named, the saving is booked as non-cash-releasing. **Before Spending Review 2027.**

The same ministry reconciles the five highest-claiming lines by asking the people whose time was freed how often they do the task, how long it takes, and whether the tool saves time. That method found 9 of 10 bots misstating their hours[12]. **In the first year.**

Each deploying department's permanent secretary runs one withheld-licence quarter: random allocation of the next tranche, a survey of the never-licensed remainder, and blind-scored observed tasks, quality scored beside time. Consultation completes before allocation, on pain of €1,000 a day[16]. **Q1 2027, before a third of headcount is covered.**

The body approving year-two funding refuses any deployed system without a published monthly recurring-user count beside its licence count, and makes every business case multiply a per-user saving by that rate. An 83 per cent adoption headline on one tool was 64 per cent weekly use[3,4]. **At the next renewal round.**

The central procuring body buys the promotional year on the $1-an-agency terms[17,18] and uses no repayable central fund; the best-documented one has booked $13.5m of savings against $1.03bn invested, and publishes no figure for cash repaid[19]. With the finance ministry it fixes the post-promotion per-user list price in writing at award and prices a costed exit the same day. It refuses any case-system business case without separate lines for record preparation, integration and a five-year run rate[23,24], and any system whose year-two budget line is not named[6]. **Before those agreements expire, 30 September 2026.**

The sponsoring ministry amends the enabling statute before go-live to carry human review, notification, and a duty to publish decisions issued, challenges lodged and outcomes changed per system per year. GDPR Article 22(2)(b) leaves that safeguard to the statute[20]; the AI Act supplies notification only from 2 December 2027[21,22]. Notice runs as policy from day one[14]. **Before go-live; deployed systems before 2 December 2027.**

The service owner funds two lines from year one: a random audit of decisions nobody challenged, sized from a sampled defect rate, reporting outside the programme, and challenge handling costed from decision volume[15]. One such review found roughly 44,000 defective determinations behind a 7 per cent appeal rate[13]. **At business-case approval.**

Policy Briefs

Short guidebooks for ministers and senior officials, published by [Your Institution]. Each issue follows one question from evidence to decision. Series editor: [Name Surname].

SOURCES CITED IN THIS BRIEF

IN ORDER OF CITATION

1 Netherlands Court of Audit, *Focus op AI*, 2024.

2 National Audit Office, *AI in government*, 2024.

3 Government Digital Service, *Copilot findings*, 2025.

4 Department for Business and Trade, *Copilot*, 2025.

5 HM Revenue and Customs, *Copilot in HMRC*, 2026.

6 Danish Digital Agency, *AI-signaturprojekter*, 2024.

7 US OMB, *Federal AI Use Case Inventory*, 2025.

8 MHCLG, *AI tool for planning decisions*, 2026.

9 DWP, *Copilot trial evaluation*, 2026.

10 National Audit Office, *Workforce planning*, 2026.

11 HM Treasury, *Government Efficiency Framework*, 2023.

12 US GSA Inspector General, *RPA savings*, 2023.

13 Michigan, *Fraud determination review*, 2017.

14 Canada, *Directive on Automated Decision-Making*, 2025.

15 UK Parliament, *Tribunal costs*, PQ 47647, 2022.

16 Capstan Avocats on TJ Nanterre n° 24/01457, 2025.

17 US General Services Administration, *OneGov*, 2025.

18 FedScoop, *OneGov AI deals are ending*, 2026.

19 US GAO, *Technology Modernization Fund*, 2026.

20 EU, *Regulation (EU) 2016/679 (GDPR)*, 2016.

21 EU, *Regulation (EU) 2024/1689 (AI Act)*, 2024.

22 EU, *Digital Omnibus on AI (2026/1744)*, 2026.

23 US GAO, *Artificial Intelligence Acquisitions*, 2026.

24 US OMB, *Memorandum M-25-22*, 2025.

25 European Commission, *SWD(2021) 84 final* (AI Act impact assessment), 2021.

HOW TO CITE

[Your Institution], *Adoption before proof*, Policy Briefs No. 05, August 2026.

CORRECTIONS & UPDATES

Evidence moves. Corrections and revised editions are published at [institution.example]; the version of record carries this issue number.

ALSO IN THIS SERIES

ISSN [0000-0000]
No. 05 · August 2026 · CC BY 4.0

[institution.example]
[briefs@institution.example]

Next in the series
No. 06 · Safety, incidents and redress