Commentary and Critique **Research Commentary**
Vol. 12 (2026) · ISSN 3006-6220 journal.pe.ac/Commentary-and-Critique

Automating Suspicion, Rationing Redress: Fraud-Risk Scoring, Debt Recovery, and the Asymmetric Burden of Error in Digital Welfare Administration

Samuel Tay[a], Nicholas Tang[a], Sophie Hadfield[a]

[a]Academy of Public Policy and Economics, London E1 0SG, United Kingdom

Corresponding author: Samuel Tay, samuel.tay@pe.ac.

Published 15 November, 2026.

DOI: https://doi.org/10.62432/APPE.CC.2026.12.57

Article webpage: https://journal.pe.ac/Commentary-and-Critique/paper/view/125683

Abstract

Welfare agencies across advanced democracies increasingly deploy automated cross-registry matching and machine-learning risk models to detect benefit overpayments, yet flagship anti-fraud initiatives have repeatedly triggered unlawful debt recovery, wrongful benefit suspensions, and multibillion-euro remediation backlogs. Comparing statutory rules, judicial rulings, and parliamentary and supreme audit inquiries across Australia's Online Compliance Intervention, the Dutch childcare allowance enforcement and restitution regime, and the United Kingdom's Universal Credit compliance architecture, this commentary argues that digital welfare failures stem from a structural cost wedge between automated suspicion and manual redress. When agencies reduce the marginal administrative cost of flagging discrepancies, suspending entitlements, or raising debts toward zero while keeping evidentiary rebuttal and appeals inside sequential, claimant-initiated manual channels, split operating and transfer budgets externalize false-positive liquidity and compliance costs onto low-income households. Retaining human caseworkers inside predisbursement advance triage prevents automated historical debt generation, yet ongoing-award case reviews, opaque risk features, and pre-verification claim suspensions preserve asymmetric error incidence, while ex post individualized restitution programs erase net fiscal gains. Rebalancing digital welfare administration requires tiered primary-record verification (using existing tax-payroll feeds before noticed third-party checks), symmetric error and claim-abandonment accounting audited by supreme audit bodies, and statutory production-model logging accessible to independent watchdogs.

Keywords: digital welfare administration; algorithmic risk scoring; administrative burden; automated debt recovery; supreme audit institutions; procedural justice

1. Introduction

Automated cross-registry matching and machine-learning fraud models now sit at the center of social security administration across advanced democracies. Ministries of social affairs and treasury departments routinely justify algorithmic compliance tools on two grounds: reducing administrative operating outlays and curbing improper benefit payments without penalizing eligible households (Henman, 2020; National Audit Office, 2024). Over the past decade, however, flagship digital anti-fraud programs have repeatedly triggered acute legal and fiscal breakdowns. Between July 2015 and November 2019, Australia's income-averaging compliance drive (expanded nationwide in July 2016 as the Online Compliance Intervention, or

© 2026 Academy of Public Policy and Economics · Open Access (CC BY 4.0).

Robodebt) raised approximately AU$1.76 billion in unlawful social security debts across roughly 433,000 to 443,000 targeted recipients (Royal Commission into the Robodebt Scheme, 2023). In the Netherlands, risk-classification profiling combined with rigid "all-or-nothing" annual clawback rules stripped tens of thousands of families of childcare allowances (*kinderopvangtoeslag*), precipitating cabinet resignation in January 2021 and a multi-year statutory remediation operation (*Hersteloperatie Toeslagen*) whose projected fiscal envelope reached roughly €8 billion by 2024 (Parlementaire ondervraging kinderopvangtoeslag, 2020; Algemene Rekenkamer, 2024). In the United Kingdom, where the Department for Work and Pensions (DWP) pairs machine-learning risk triage on Universal Credit Advances with human-staffed Targeted Case Reviews of ongoing awards, supreme audit scrutiny in 2025 identified persistent opacity in risk features, incomplete fairness linkage, and continuing tension between fraud containment and claimant protection (Department for Work and Pensions, 2025; National Audit Office, 2025).

Existing public administration scholarship illuminates major facets of these episodes yet leaves their shared fiscal-procedural engine under-specified. Foundational work on system-level and infrastructure-level bureaucracies shows how digital architectures displace street-level discretion into software code and inter-agency data pipelines (Bovens & Zouridis, 2002; Bullock, 2019; Peeters & Widlak, 2023). Legal and social policy critiques of individual scandals document statutory departures, punitive political rhetoric, and heavy compliance burdens imposed on disadvantaged claimants (Moynihan et al., 2015; Carney, 2019; Herd & Moynihan, 2019; Whiteford, 2021). Meanwhile, regulatory governance studies frequently prescribe "human-in-the-loop" caseworker sign-off as the primary safeguard against automated error (Yeung, 2018; Meijer et al., 2021). Treating these failures primarily as isolated coding flaws, rogue ministerial episodes, or absences of human sign-off misses a deeper structural regularity: why rational, budget-constrained welfare agencies repeatedly adopt digital workflows that shift the cost of administrative error onto low-income households, and why inserting human reviewers into the referral queue only partially attenuates that shift.

This commentary argues that recurring breakdowns in digital welfare enforcement arise from an asymmetric administrative cost wedge between automated suspicion and rationed manual redress, reinforced by split public budgeting rules. When cross-agency batch matching or machine-learning risk scoring lowers an agency's marginal staffing cost of flagging an account, withholding a payment, or issuing a discrepancy notice toward zero, while keeping evidentiary verification, historical income reconstruction, and formal appeals inside sequential, claimant-initiated manual channels, departmental budget structures reward over-flagging. Because primary verification consumes capped departmental operating budgets whereas gross detected overpayments and claimant non-engagement drops register as savings in demand-led benefit transfer accounts, false-positive liquidity shocks and document-retrieval burdens fall off-budget onto claimants (Christensen et al., 2020; Redden et al., 2020). Comparing statutory provisions, judicial rulings, and parliamentary and supreme audit inquiries across Australia, the Netherlands, and the United Kingdom through November 2026 yields two institutional conclusions. First, disaggregating the British model shows that while human-reviewed triage on pre-disbursement advance loans avoids unlawful historical debt creation, ongoing-award case reviews, opaque risk scores, and non-response claim suspensions preserve asymmetric error incidence (Alon-Barkat & Busuioc, 2023; National Audit Office, 2025). Second, when front-end automation and cliff-edge clawback rules culminate in systemic legal invalidation, ex post individualized restitution programs consume multibillion-euro administrative and fiscal outlays that erase intended compliance savings and generate secondary financial-regularity failures (Algemene Rekenkamer, 2023; Algemene Rekenkamer, 2025).

Section 2 develops the analytical framework linking asymmetrical verification costs, substantive sanction proportionality, and public financial management incentives, defining an explicit five-regime institutional rubric. Section 3 compares proxy-triggered enforcement and ex post remediation bottlenecks in Australia and the Netherlands. Section 4 disaggregates the United Kingdom's Universal Credit Advances

machine-learning triage from ongoing Targeted Case Reviews. Section 5 evaluates rival explanations centered on fiscal urgency, predictive model precision, and non-payroll eligibility domains. Section 6 formulates three governance safeguards reconciled with European and Commonwealth data-protection law.

2. The Asymmetric Administrative Cost Wedge in Digital Welfare

2.1. Zero-Marginal-Cost Suspicion, Sanction Disproportionality, and Manual Redress

In paper-era welfare administration, initiating an overpayment investigation imposed a tangible resource constraint on the enforcing agency. Before suspending an established entitlement or raising a recoverable debt, compliance staff had to issue formal statutory notices to employers or financial institutions, reconcile mismatched reporting intervals, and establish period-specific lack of entitlement (Commonwealth Ombudsman, 2017; Parliament of Australia, 1999; Whiteford, 2021). Because investigative staff hours were scarce, administrative cost acted as an implicit brake on speculative enforcement: agencies concentrated manual audits where expected recoveries justified investigative effort.

Digital compliance architectures sever this proportionality at the front end while preserving manual friction at the back end. Automated batch matching across tax, residency, and benefit registries, or machine-learning risk models trained on past investigations, enable central analytics teams to screen millions of claims at near-zero marginal staffing cost per additional flag (Bovens & Zouridis, 2002; Peeters & Widlak, 2023). Yet proxy mismatches and probabilistic risk scores rarely establish lawful overpayment liability on their own. Annual tax summaries do not align with fortnightly benefit assessment windows; household childcare hours vary across school terms; and statistical risk features correlate with lawful life transitions such as casualized employment, address changes, or relationship breakdown (Carney, 2019; National Audit Office, 2025). Determining whether a flagged claim reflects genuine fraud, reporting lag, or proxy measurement noise still requires individualized inspection of primary records.

The decisive institutional choices lie in *who* must perform that costly reconciliation, *what happens to benefit payments in the interim*, and *how severely minor documentary gaps are sanctioned*. As shown in Figure 1 (Panel a), digital enforcement regimes diverge along two procedural pathways. Under Pathway I (proxy-triggered withholding or clawback), the agency treats an automated discrepancy flag or risk-score breach as sufficient to suspend payments, demand repayment, or invert the evidentiary onus onto the recipient. Where procedural inversion is further paired with cliff-edge substantive forfeiture rules, such as revoking an entire annual family subsidy over a minor invoice defect or missing 2% co-payment (Raad van State, 2019; Parlementaire ondervraging kinderopvangtoeslag, 2020), front-end automation converts routine paperwork friction into household insolvency. To contest the presumption, claimants must navigate online portals, retrieve multi-year payslips or childcare receipts from former employers or providers, and queue for rationed manual review staff (Moynihan et al., 2015; Herd & Moynihan, 2019). Under Pathway II (tiered verify-before-withholding), the automated score operates solely as a logged internal triage signal: ongoing subsistence entitlements continue while the agency checks existing period-matched tax feeds first and observes statutory notice safeguards before external inquiries.

2.2. Split Public Budgeting and Liquidity Incidence

Public financial management rules explain why ministries repeatedly gravitate toward Pathway I despite its legal hazards. Across Westminster and continental fiscal systems, statutory benefit transfers flow through demand-led entitlement accounts (such as Annually Managed Expenditure in the United Kingdom or statutory benefit chapters in Australia and the Netherlands), whereas compliance caseworkers and verification infrastructure are funded out of capped departmental operating appropriations (such as UK Departmental Expenditure Limits) (Whiteford, 2021; Department for Work and Pensions, 2025). Simulta-

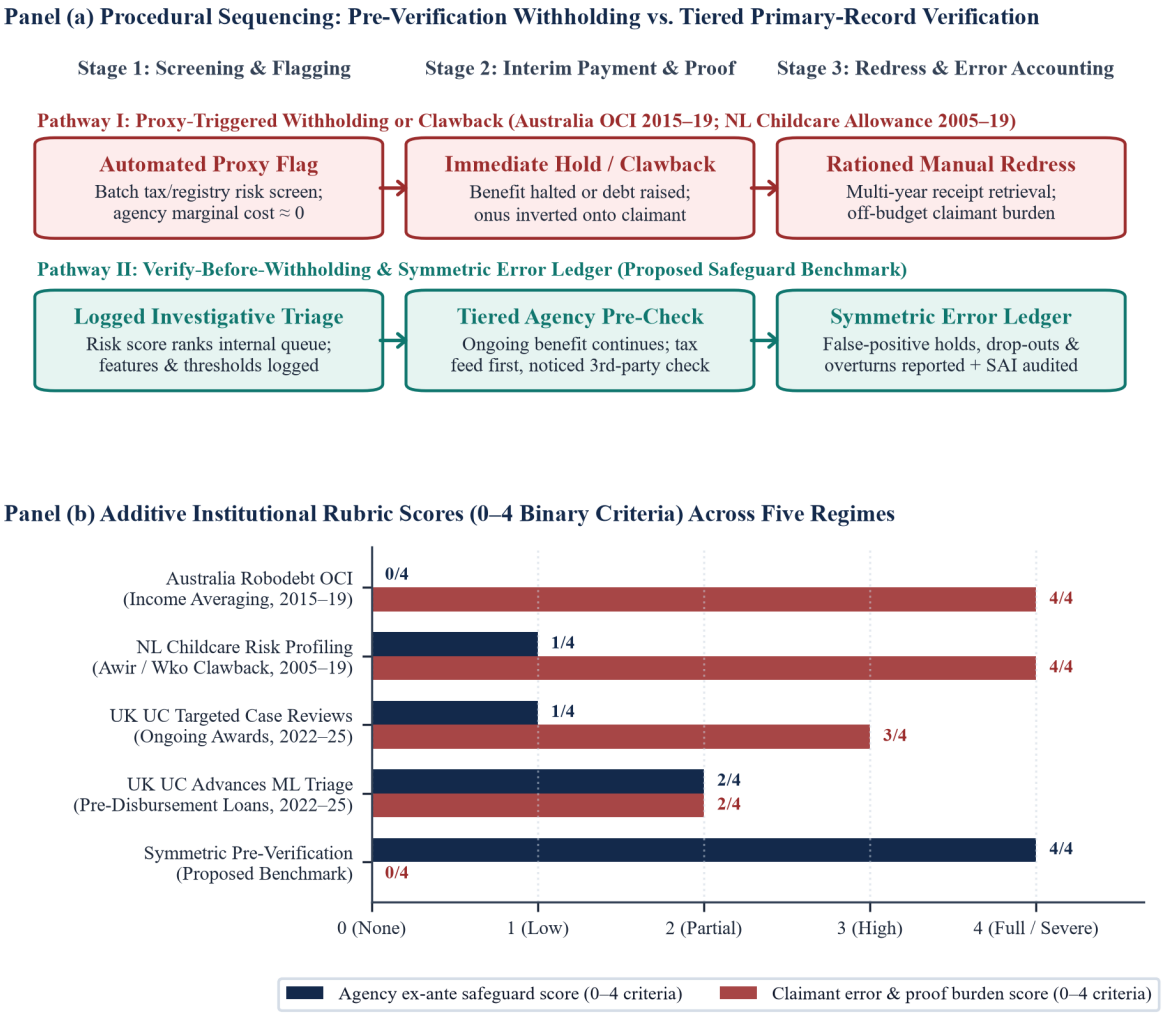

Figure 1. The asymmetric administrative cost wedge and additive institutional rubric scores across five welfare compliance regimes.

Notes. Panel a contrasts procedural sequencing under proxy-triggered pre-verification withholding or clawback (Pathway I) against the proposed safeguard benchmark combining logged triage, tiered agency primary-record verification, and symmetric error accounting (Pathway II). Panel b reports exact integer sums (0 to 4) across five regimes on two additive four-criterion binary rubrics. Agency ex-ante safeguard score sums four binary indicators (1 = yes, 0 = no): (i) mandatory human caseworker review before adverse action; (ii) uninterrupted ongoing subsistence entitlement during initial screening; (iii) affirmative agency duty to verify primary records prior to shifting proof onto the claimant; and (iv) proportional interim adjustment paired with externally audited error/abandonment accounting. Claimant error and proof burden score sums four binary hazards (1 = present, 0 = absent): (i) evidentiary onus shifted onto claimant to supply multi-month or multi-year primary documents; (ii) opaque risk features or unverified proxy averages withheld from claimant scrutiny; (iii) immediate liquidity loss via active benefit suspension or pre-adjudication tax-refund garnishment; and (iv) retrospective multi-year debt imposition, 100% cliff-edge annual forfeiture, or inter-agency blacklist propagation.

neously, treasury budget agreements evaluate agency executives on gross compliance savings targets and headline monetary overpayment rates. Within this split budget architecture, requiring compliance staff to corroborate primary records ex ante consumes scarce capped operating funds, whereas shifting verification onto claimants conserves staff hours while booking gross recoveries and non-engagement claim terminations as transfer-account savings. False positives (eligible recipients subjected to erroneous payment holds, protracted document demands, or wrongful debt recovery who abandon claims out of exhaustion) impose zero charge against either departmental account (Carney, 2019; Christensen et al., 2020).

This budgetary asymmetry compounds a stark disparity in household liquidity tolerance. Delaying a £600 monthly subsistence payment or clawing back a disputed €4,000 childcare advance is trivial for a national exchequer, yet triggers immediate rent arrears, utility shutoffs, and penalty cascades for liquidity-constrained families (Herd & Moynihan, 2019; European Commission for Democracy through Law, 2021). Cognitive and compliance burdens further stratify outcomes inside the redress queue: claimants facing housing instability, disability, language barriers, or irregular gig work exhibit the highest likelihood of triggering atypical-pattern flags while possessing the fewest archival and legal resources to pursue multi-stage administrative appeals (Christensen et al., 2020; Peeters & Widlak, 2023). Evaluating all five institutional designs on the explicit four-criterion binary rubrics in Figure 1 (Panel b) illustrates the resulting inverse pattern: Australia's OCI (0/4 safeguards; 4/4 claimant hazards) and pre-2019 Dutch childcare enforcement (1/4; 4/4) concentrate error burdens entirely on households, whereas Britain's two Universal Credit compliance tracks occupy distinct intermediate positions (1/4 vs. 3/4 for ongoing Targeted Case Reviews; 2/4 vs. 2/4 for pre-disbursement Advances triage).

3. Proxy Automation and Evidentiary Inversion: Australia and the Netherlands

Table 1 compares the statutory authorities, screening inputs, human-review boundaries, interim payment rules, and inquiry and audit records across Australia, the Netherlands, and the two disaggregated United Kingdom Universal Credit instruments.

Table 1. Institutional architecture of automated compliance screening and redress across four welfare compliance mechanisms in three jurisdictions.

Dimension	Australia: Online Compliance Intervention (2015–2019)	Netherlands: Childcare Allowance Profiling & Restitution (2005–2025)	UK Universal Credit: Advances ML Triage & Targeted Case Reviews (2022–2025)
Primary statutory framework	*Social Security Act 1991* (Cth) ss. 1064, 1223(1), 1228B; *Social Security (Admin.) Act 1999* ss. 63, 80, 192–196	*Wet kinderopvang* (Wko, Arts. 1.6–1.7); *Awir* (Arts. 16, 26); *Awb* Art. 3:4(2); *Wht* (*Stb.* 2022, 433)	*Social Security Admin. Act 1992* s. 71ZB; *UC Regs.* SI 2013/383 (Advances) & SI 2013/380 regs. 37–38, 44–47 (TCR); 2025 c. 28
Automated screening input	Batch match of annual ATO PAYG employer totals against fortnightly Centrelink declarations	*Risicoclassificatiemodel* (deployed 2013) monthly risk scores & cross-registry flags	*Advances*: ML risk score on lump-sum loan claims; *TCR*: risk/analytics selection of active UC awards
Automated vs. human boundary	Automated income averaging & debt raising when portal letter went unanswered	High risk scores guided manual audits; rigid administrative & pre-2019 judicial full clawback	*Advances*: caseworker check required (1-day median); *TCR*: human review staff demand multi-month bank statements
Interim benefit & debt rule	Averaged debt raised; tax refunds garnished & 10% s. 1228B fee added	Immediate stop & 100% annual subsidy clawback (*nihilstelling*) for minor gaps	*Advances*: new loan held pending ID proof (ongoing UC unaffected); *TCR*: active UC suspended/closed upon non-response
Evidentiary burden allocation	Inverted onto recipient to upload years-old fortnightly payslips via portal	Inverted onto parents to prove complete annual invoices & 100% fee co-payments	Applicant/claimant must upload ID or bank statements; risk features withheld from public scrutiny
Primary inquiry & SAI audit record	Ombudsman (02/2017); Royal Commission (July 2023); APSC Taskforce (Sept 2024)	*Tweede Kamer* (*Ongekend onrecht*, Dec 2020); *Algemene Rekenkamer* (2023–2025)	DWP Annual Accounts HC 995 (July 2025); NAO SAI Reports HC 612 (2024) & HC 1336 (Oct 2025)

Notes. Compiled from Parliament of Australia (1991), Parliament of Australia (1999), Commonwealth Ombudsman (2017), Raad van State (2019), Parlementaire ondervraging kinderopvangtoeslag (2020), Staten-Generaal der Nederlanden (2022), Royal Commission into the Robodebt Scheme (2023), Australian Public Service Commission (2024), Algemene Rekenkamer (2024), Algemene Rekenkamer (2025), Department for Work and Pensions (2025), National Audit Office (2025), and Parliament of the United Kingdom (2025). SAI denotes Supreme Audit Institution.

3.1. Australia's Online Compliance Intervention: Statutory Bypass and Fiscal Reversal

Under section 1064 and section 1223(1) of Australia's *Social Security Act 1991* (Cth), lawful overpayment calculation turns strictly on actual income earned within each fortnightly entitlement period: a student or casual worker earning nothing during university semesters lawfully qualifies for full fortnightly support even if higher wages later in the financial year raise annual income (Parliament of Australia, 1991; Carney, 2019). Prior to 2015, the Department of Human Services (DHS) treated annual Australian Taxation Office (ATO) pay-as-you-go summaries solely as an investigative lead; when recipients lacked historical payslips, compliance officers exercised statutory information-gathering powers under sections 192, 195, and 196 of the *Social Security (Administration) Act 1999* (Cth) to obtain fortnightly payroll records directly from employers, completing roughly 20,000 manual reviews per year (Parliament of Australia, 1999; Commonwealth Ombudsman, 2017).

Driven by a 2015–16 federal Budget compliance measure projecting AU$1.70 billion in fiscal savings (*Strengthening the Integrity of the Welfare System*), DHS piloted removing manual employer checks in July 2015 and launched the full self-service Online Compliance Intervention portal nationwide in July 2016 (Whiteford, 2021; Royal Commission into the Robodebt Scheme, 2023). Bypassing formal s. 192 employer notices, the system mailed automated discrepancy letters directing current and former recipients onto an online portal. Where recipients did not respond, often because letters went to outdated addresses, or could not retrieve fortnightly payslips dating back up to seven years, the software divided annual ATO earnings into equal fortnightly averages, raised a purported s. 1223(1) debt, added an unlawful 10% s. 1228B recovery fee, garnished tax refunds, and engaged private collection agencies (Royal Commission into the Robodebt Scheme, 2023).

Crucial to comparative accountability is *which* institutions failed or succeeded in halting the scheme. Unlike the British and Dutch supreme audit offices examined below, the Australian National Audit Office did not conduct an early performance audit stopping unlawful averaging; meanwhile, senior DHS and social-services executives suppressed internal 2014 legal advice warning of statutory invalidity and supplied incomplete representations that blunted the Commonwealth Ombudsman (2017) investigation (Royal Commission into the Robodebt Scheme, 2023). Only Federal Court test litigation in *Amato v Commonwealth* (27 November 2019) forced concession, followed by the *Prygodicz v Commonwealth (No 2)* ([2021] FCA 634) class-action settlement and the July 2023 Royal Commission report. Across 2015–2019, the scheme raised AU$1.76 billion in unlawful debts (>470,000 distinct debts across ~433,000 to 443,000 targeted individuals), compelling the Commonwealth to refund AU$746 million in cash repayments to ~381,000 individuals, zero out remaining balances, and pay AU$112 million in court settlement costs, erasing the projected AU$1.70 billion budget saving entirely (Figure 2, Panel a; Table 2). Although the Australian Government (2023) accepted or accepted in principle 56 of 57 Royal Commission recommendations (declining full repeal of the Cabinet-document Freedom of Information exemption) and the Australian Public Service Commission (2024) found 97 Code of Conduct breaches across 12 of 16 referred officials, section 15 of the *Public Service Act 1999* (Cth) precluded disciplinary sanctions against departed agency heads.

3.2. The Dutch Childcare Allowance Crisis and the Fiscal-Regularity Trap of Ex Post Restitution

Whereas Robodebt relied on deterministic arithmetic averaging, the Dutch childcare allowance crisis (*Toeslagenaffaire*) illustrates how algorithmic risk profiling interacts with disproportionate statutory

forfeiture rules and inter-agency data sharing (Peeters & Widlak, 2023). Under the Childcare Act (*Wet kinderopvang*, Wko, enacted 2005) and the General Act on Income-Dependent Schemes (*Awir*, 2005), Dutch working parents received monthly provisional childcare advances reconciled after year-end (Parlementaire ondervraging kinderopvangtoeslag, 2020). Following political mobilization over cross-border fraud in 2013, the Tax and Customs Administration (*Belastingdienst/Toeslagen*) deployed a monthly self-adjusting risk-classification model (*risicoclassificatiemodel*) incorporating indicators such as non-Dutch or dual nationality, single parenthood, and prior estimates to select families for intensive verification, paired with internal blacklists in the Fraud Signaling Facility (*Fraude Signalering Voorziening*), for which the Dutch Data Protection Authority (*Autoriteit Persoonsgegevens*) later imposed GDPR fines of €2.75 million and €3.7 million (European Commission for Democracy through Law, 2021; Peeters & Widlak, 2023).

Algorithmic targeting proved ruinous because it fed into two rigid legal doctrines upheld for years by the Administrative Jurisdiction Division of the Council of State (*Raad van State*) until its companion reversals of 23 October 2019 (Raad van State, 2019). Under Wko Articles 1.6 and 1.7 (reversed in ECLI:NL:RVS:2019:3535), failing to document timely payment of even 2% of annual invoiced day-care fees triggered complete zeroing (*nihilstelling*) of the *entire* annual subsidy; under Awir Articles 16 and 26 (reversed in ECLI:NL:RVS:2019:3536 via Awb Article 3:4(2) proportionality review), the agency enforced mandatory 100% clawback (*terugvordering*) of tens of thousands of euros without individualized hardship relief (Raad van State, 2019; European Commission for Democracy through Law, 2021). Infrastructure-level data sharing propagated unverified fraud tags across municipal debt, housing, and child-welfare agencies while front-line staff could neither view nor correct the underlying flag (Peeters & Widlak, 2023).

What followed the December 2020 parliamentary inquiry report *Ongekend onrecht* demonstrates the severe public-management trap of ex post individualized remediation. Codified under the Child-care Allowance Restitution Act (*Wet hersteloperatie toeslagen*, Wht, *Stb.* 2022, 433; Staten-Generaal der Nederlanden, 2022), remediation required adjudicating heterogeneous multi-year life damages across >35,000 recognized parents (screened from >68,000 applicants). Successive accountability audits by the Netherlands Court of Audit (Algemene Rekenkamer, 2023; Algemene Rekenkamer, 2024; Algemene Rekenkamer, 2025) document that total projected multi-year outlays climbed from an initial €500 million reserve in 2020 to roughly €8 billion by 2024 (~€5.2 billion in direct compensation plus public and private debt write-offs, municipal support, and heavy contractor staffing; Figure 2, Panel b; Table 2). Politically mandated leniency (*ruimhartigheid*) under the rapid €30,000 flat-rate *Catshuisregeling* produced both horizontal inequities and formal financial-regularity errors (*rechtmatigheidsfouten* and *onzekerheden* totaling ~13%, or €113.5 million, of audited 2023 restitution outlays), while comprehensive assessments (*integrale beoordelingen*) and supplementary actual-damage claims (*aanvullende schadeverzoeken*) remained backed up toward 2027, straining IT and personnel capacity across the tax administration (Algemene Rekenkamer, 2024; Algemene Rekenkamer, 2025). Batch-screening millions of returns costs minimal staff time ex ante; reconstructing counterfactual household damages years later degrades both fiscal efficiency and budgetary regularity.

Table 2. Official inquiry and Supreme Audit Institution indicators of compliance scale, reversal, and remediation bottlenecks across the three jurisdictions.

Jurisdiction, accounting basis & metric	Verified official magnitude	Reference window	Primary official source & denominator boundary
Australia OCI (stock): Targeted cohort & debts	~433k–443k people (>470k debts)	Jul 2015 – Nov 2019	Royal Commission (2023, Vol. 1); ~381k individuals received cash refunds

(continued)

Table 2 (continued)

Jurisdiction, accounting basis & metric	Verified official magnitude	Reference window	Primary official source & denominator boundary
Australia OCI (stock): Gross unlawful debts vs. target	AU$1.76B raised (vs. AU$1.70B target)	Jul 2015 – Nov 2019	Royal Commission (2023); erases 2015–16 Budget projected savings
Australia OCI (stock): Cash refunded & settlement	AU$746M refunded + AU$112M settlement	2020 – 2023	Royal Commission (2023); cash refunds plus *Prygodicz* [2021] FCA 634
Australia APS inquiry: Code of Conduct findings	97 breaches across 12 of 16 officials	Sept 2024	APSC (2024); s. 15 *Public Service Act 1999* barred post-separation fines
Netherlands *Toeslagen* (cohort): Recognized parents	>35,000 parents (>68,000 applicants)	2020 – 2025	*Algemene Rekenkamer* (2024, 2025); recognized *gedupeerden* under *Wht*
Netherlands *Hersteloperatie* (multi-yr envelope)	~€8.0B projected (~€5.2B direct comp.)	2020 – 2027 proj.	*Algemene Rekenkamer* (2024, 2025); up from €0.5B initial 2020 reserve
Netherlands restitution regularity & flat payout	€30,000 flat rate; ~13% (€113.5M) 2023 errors	FY 2021 – 2024	*Algemene Rekenkamer* (2023, 2024, 2025); *onrechtmatigheden* & 2027 backlog
UK DWP (annual flow): All-benefit overpayments	£9.5B (3.3% of total benefit outlay)	FY 2024–25	DWP HC 995 (July 2025); down from £9.7B (3.6%) across all DWP benefits
UK Universal Credit (annual flow): Monetary & rate	£6.35B (9.7% of UC expenditure)	FY 2024–25	NAO HC 1336 (Oct 2025); vs. £6.41B (12.4%) in 2023–24 amid UC migration
UK UC Advances (annual flow): ML-screened loans	1.4M advances (£0.80B; £0–£60M fraud/err.)	FY 2024–25	DWP HC 995 & NAO HC 1336 (2025); 1-day median human triage delay

Notes. All indicators reflect official inquiry, statutory, and Supreme Audit Institution reports published on or before 13 November 2026: Royal Commission into the Robodebt Scheme (2023), Australian Public Service Commission (2024), Algemene Rekenkamer (2023), Algemene Rekenkamer (2024), Algemene Rekenkamer (2025), Department for Work and Pensions (2025), and National Audit Office (2025). Stocks (AU), multi-year remediation budget envelopes (NL), and single-year gross program flows (UK) represent distinct public-accounting categories and are not summed across countries.

4. Disaggregating Human-in-the-Loop Governance: The United Kingdom's Two Universal Credit Compliance Tracks

4.1. Pre-Disbursement UC Advances Triage Versus Ongoing Targeted Case Reviews

European and Commonwealth digital-government standards increasingly invoke "human-in-the-loop" review as the primary institutional remedy against Robodebt-style illegality (Yeung, 2018; National Audit Office, 2024). Evaluating the United Kingdom's Department for Work and Pensions requires disaggregating two legally and operationally distinct Universal Credit compliance mechanisms that are often conflated in public debate: (1) pre-disbursement machine-learning risk triage applied to lump-sum *Universal Credit Advances* under the *Universal Credit (Payments on Account) Regulations 2013* (SI 2013/383), and (2) retrospective and ongoing *Targeted Case Reviews* (TCR) of active awards governed by suspension and termination powers under the *Universal Credit, Personal Independence Payment, Jobseeker's Allowance and Employment and Support Allowance (Claims and Payments) Regulations 2013* (SI 2013/380, regs. 37–38, 44–47) alongside section 71ZB of the *Social Security Administration Act 1992* (Department for Work and Pensions, 2025; National Audit Office, 2025).

Official accounts for financial year 2024–25 (**HC 995** and **HC 1336**) clarify both what British human-triage governance achieves and what headline percentage rates obscure. Across DWP as a whole, total benefit overpayments stood at £9.5 billion (3.3% of total benefit expenditure, down from £9.7 billion or 3.6% in 2023–24). Within Universal Credit specifically, the headline overpayment *rate* fell from 12.4% in 2023–24 to 9.7% in 2024–25, yet **absolute monetary UC overpayments barely budged, edging from £6,410 million (£6.41 billion) to £6,350 million (£6.35 billion)** because rapid "Move to UC" migration of legacy benefit claimants expanded the total Universal Credit expenditure denominator (Department for Work and Pensions, 2025; National Audit Office, 2025; Figure 2, Panel c; Table 2). Within the Ad-

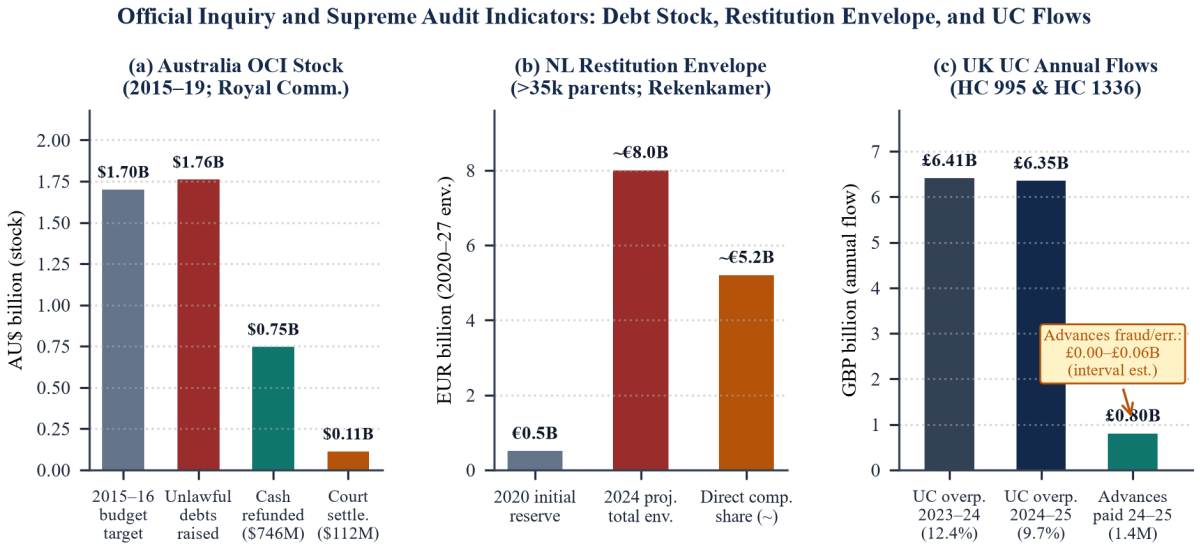

Figure 2. Official inquiry and supreme audit indicators: Australian unlawful debt stock, Dutch multi-year restitution envelope, and UK Universal Credit annual flows.

Notes. Panels distinguish three non-identical accounting bases across jurisdictions. Panel a reports cumulative 2015–2019 Australian OCI stock magnitudes: the initial 2015–16 Budget savings target (AU$1.70B), gross unlawful averaged debts raised (AU$1.76B), cash repayments refunded (AU$0.75B [AU$746M]), and Federal Court settlement outlays (AU$0.11B [AU$112M]), excluding internal agency staffing costs (Royal Commission into the Robodebt Scheme, 2023). Panel b reports the multi-year 2020–2027 Dutch *Hersteloperatie Toeslagen* total budget envelope (~€8.0B projected by 2024 vs. €0.5B initial 2020 reserve, including ~€5.2B in direct compensation alongside debt assumptions and administrative execution outlays) (Algemene Rekenkamer, 2024; Algemene Rekenkamer, 2025). Panel c reports single-year UK Universal Credit annual monetary flows for FY 2023–24 and FY 2024–25: estimated UC overpayments (£6.41B [12.4%] and £6.35B [9.7%]) and FY 2024–25 UC Advances disbursed (£0.80B across 1.4M advances, with estimated advances fraud and error spanning the interval £0.00–£0.06B) (Department for Work and Pensions, 2025; National Audit Office, 2025).

vances sub-stream (introduced to counter pandemic identity hijacking by organized crime groups), DWP disbursed 1.4 million repayable advances worth £0.80 billion in 2024–25, holding estimated residual advances fraud and error within an interval of £0 to £60 million. By policy, DWP prohibits purely automated refusal: elevated machine-learning scores route advance applications to human caseworkers with a reported median referral delay of roughly one day, without generating retrospective tax-averaged debts (Department for Work and Pensions, 2025; National Audit Office, 2025).

4.2. Why Human Triage Leaves Error Asymmetry Intact Across Advances and Ongoing Reviews

Closer scrutiny of **HC 995**, **HC 1336**, and behavioral public administration evidence shows why human sign-off only partially narrows the cost wedge, leaving UC Advances ML Triage at 2/4 and UC Targeted Case Reviews at 3/4 on claimant burden (Figure 1, Panel b).

First, under high throughput quotas and opaque feature weights, human caseworkers exhibit systematic *automation bias* and *selective adherence* (Alon-Barkat & Busuioc, 2023; Alon-Barkat & Busuioc, 2024). When a machine-learning screen labels an advance claim "high risk" without exposing feature contributions or false-positive base rates to the caseworker, overriding the alert exposes the reviewer to audit blame if fraud later materializes, whereas demanding extra identity or tenancy uploads imposes zero departmental cost (Bullock, 2019; Meijer et al., 2021). Even on pre-disbursement advances, applicants lacking immediate digital upload capacity, such as individuals fleeing domestic abuse or in unstable

temporary housing, forfeit urgent bridge liquidity without any adverse decision formally entering appeal statistics (Herd & Moynihan, 2019; Redden et al., 2020).

Second, ongoing-award Targeted Case Reviews reproduce direct subsistence interruption risk. Unlike pre-disbursement advance checks, TCR teams instruct active Universal Credit households to upload multi-month bank statements and evidence within strict deadlines under SI 2013/380 (regs. 37–38, 44–47). If a claimant fails to respond in their online journal, DWP suspends and subsequently terminates the entire ongoing award, booking non-engagement exits as multi-billion-pound AME fraud-and-error savings without systematically distinguishing ineligible exits from cognitively or digitally overwhelmed eligible households (Christensen et al., 2020; National Audit Office, 2025).

Third, public opacity limits external disparate-impact verification. DWP withholds risk features and thresholds from claimants and external researchers to prevent evasion, while publishing subgroup referral-rate ratios in **HC 995** (Department for Work and Pensions, 2025). Reviewing these arrangements in **HC 1336**, the National Audit Office (2025) identified incomplete administrative data linkage across protected characteristics and emphasized that monitoring referral disparities alone fails to capture downstream claim abandonment, delayed-payment hardship, or unappealed terminations.

5. Rival Explanations, Non-Payroll Eligibility Domains, and Surveillance Bounds

Three substantive counter-arguments test the operational limits of our framework.

First, defenders of pre-verification holds point to £6.35 billion in annual Universal Credit overpayments (£9.5 billion across all DWP benefits) and pandemic impersonation attacks as proof that immediate withholding is fiscally indispensable (Department for Work and Pensions, 2025). That objection highlights why institutional design must separate *unverified first-time credit advances* from *established subsistence entitlements*. Brief one-day human identity triage before disbursing a newly requested lump-sum loan (UC Advances) protects against irreversible identity-theft payouts without interrupting an established household award, provided non-digital fallback channels exist (National Audit Office, 2025). Conversely, suspending ongoing Australian income support, Dutch family allowances, or established Universal Credit awards prior to agency verification inflicts immediate housing and subsistence harm while risking massive downstream litigation and restitution losses (Royal Commission into the Robodebt Scheme, 2023; Algemene Rekenkamer, 2024).

Second, techno-optimist accounts attribute Robodebt and *Toeslagenaffaire* to crude early rules and predict that modern machine-learning classifiers will naturally eliminate false-positive burdens (Henman, 2020). Base-rate arithmetic refutes that claim. Even if ex ante application fraud prevalence stood near 5% to 8% (noting that DWP's £0–£60 million estimate on £0.80 billion of UC Advances, or <=7.5%, measures *post-screening* residual fraud and error rather than ex ante application prevalence), screening millions of claims with a model achieving 90% sensitivity and 90% specificity still generates more false-positive referrals than true fraud hits. Agencies also calibrate cutoffs in response to budget incentives: so long as treasury scorecards reward gross detected overpayments while pricing claimant compliance friction at zero, higher predictive accuracy tempts managers to lower referral cutoffs and expand review volume (Bullock, 2019; Alon-Barkat & Busuioc, 2024).

Third, public-law and operational critics rightly warn that shifting verification from claimants onto unfettered state data harvesting can trigger severe privacy violations, while non-payroll eligibility facts (undeclared self-employment, household cohabitation, or unlinked capital above £16,000) do not appear in employer tax feeds (National Audit Office, 2025). In *NJCM et al. v. Staat der Nederlanden* (ECLI:NL:RBDHA:2020:1878), the District Court of The Hague struck down the Dutch *SyRI* cross-registry welfare risk system under Article 8(2) of the European Convention on Human Rights for lacking verifiable necessity and data minimization (Rechtbank Den Haag, 2020). Parallel civil-liberties objections greeted the United

Kingdom's *Public Authorities (Fraud, Error and Recovery) Act 2025* (2025 c. 28, Royal Assent 2 December 2025), whose Eligibility Verification Notices compel banks to run automated bulk sweeps across benefit recipients' accounts for capital or overseas-stay indicators (Parliament of the United Kingdom, 2025), while compulsory employer notices under Australia's *1999 Administration Act* (ss. 192–196) risk stigmatizing casual workers if issued without prior claimant notice. Accordingly, agency primary-record verification must be legally tiered rather than authorizing suspicionless financial surveillance.

6. Rebalancing Administrative Justice: Three Governance Safeguards

Correcting the incentive distortions documented across Australian, Dutch, and British inquiries requires binding procedural, budgetary, and audit reforms rather than opaque human-in-the-loop sign-off alone.

First, legislatures should enact a *tiered verify-before-withholding rule paired with a staged non-response protocol* for ongoing subsistence entitlements. For wage-earning discrepancies, agencies must first corroborate period-matched statutory tax-payroll feeds already collected for tax administration (such as UK HMRC Real Time Information and post-2019 Australian ATO Single Touch Payroll) before burdening claimants. Where eligibility turns on non-payroll dimensions (capital, self-employment, or childcare invoices) requiring external bank or third-party inquiries, statutes must respect GDPR Article 5(1)(c) data minimization and Article 8 ECHR proportionality (Rechtbank Den Haag, 2020): prohibiting bulk suspicionless bank sweeps (2025 c. 28), requiring individualized reasonable grounds, and giving the claimant a non-punitive window to supply records before an employer or third party is contacted (Parliament of Australia, 1999). If an ongoing recipient fails to answer a digital review request, agencies should be barred from automatic deadline termination; instead, statute should require verified multi-channel outreach (postal, SMS, telephone, and vulnerability-marker screening), followed by a capped provisional benefit taper, permitting full suspension only as a last resort after documented multi-channel non-cooperation and senior caseworker sign-off, with full retroactive restoration upon subsequent verification and proportional *pro rata* recalculation rather than 100% annual forfeiture (Raad van State, 2019).

Second, parliamentary accounts and treasury settlements must establish *DEL-protected symmetric error accounting backed by random supreme audit sampling*. Alongside gross overpayment estimates, welfare departments should be required by statute to report audited false-positive referral rates (reviewed claims confirmed compliant), total claimant-days of wrongful payment suspension, post-referral claim-abandonment rates, and mandatory-reconsideration and tribunal overturn rates (Christensen et al., 2020; National Audit Office, 2025). To prevent automatic statutory compensation for wrongful holds from cannibalizing capped front-line caseworker budgets (DEL) and to deter Goodhart gaming (such as reclassifying compliant cases as "non-engagement" closures), compensation payouts should draw from a dedicated Treasury surcharge line penalizing executive performance scorecards while Supreme Audit Institutions audit random samples of closed non-engagement terminations and unappealed adverse decisions each year.

Third, legislatures must pair public anti-gaming confidentiality with *mandatory production-model logging, regulator parity, and pierced executive secrecy*. Supreme Audit Institutions already hold broad statutory document-access powers under section 8 of the UK *National Audit Act 1983*, sections 32–33 of Australia's *Auditor-General Act 1997* (Cth), and Articles 87–91 of the Dutch *Comptabiliteitswet 2016*. Across our three cases, external accountability faltered for three specific legal reasons: agencies failed to log protected-characteristic and post-referral abandonment data in auditable form (National Audit Office, 2025), administrative ombudsmen and statutory equality/privacy regulators lacked model-inspection parity with supreme auditors, and executives withheld adverse legal opinions behind Cabinet-document and legal-professional-privilege claims (Parlementaire ondervraging kinderopvangtoeslag, 2020; Royal Commission into the Robodebt Scheme, 2023). Requiring agencies to maintain immutable production logs

of model features, threshold changes, and longitudinal claimant drop-out outcomes, extending technical inspection rights to ombudsmen and statutory equality/data-protection watchdogs, and barring Cabinet-secrecy withholding of legal validity opinions from parliamentary and supreme auditors provide the necessary institutional foundation to align digital compliance with lawful entitlement delivery.

Statements

Data availability. This commentary relies on publicly accessible statutes, judicial decisions, parliamentary inquiry reports, and Supreme Audit Institution evaluations published on or before 13 November 2026 by the Commonwealth of Australia (Federal Register of Legislation, Royal Commission into the Robodebt Scheme, Commonwealth Ombudsman, and Australian Public Service Commission), the Kingdom of the Netherlands (Tweede Kamer, Rechtspraak.nl, Staatsblad, and Algemene Rekenkamer), the Council of Europe Venice Commission, and the United Kingdom Parliament, National Audit Office, and Department for Work and Pensions (HC 612, HC 995, HC 1336, and 2025 c. 28). Full bibliographic identifiers and archival URLs are listed in the references.

Author contributions. Samuel Tay conceived the comparative institutional framework, analyzed the Australian and British statutory and audit records, and wrote the initial manuscript draft. Nicholas Tang examined the Dutch parliamentary, judicial, and Court of Audit evaluations and developed the administrative burden synthesis. Sophie Hadfield audited all statutory pinpoints, data-protection boundaries, and comparative governance tables. All authors revised and approved the final text.

AI declaration. Bounded assistance from OpenAI Codex was used solely for cross-checking Dutch-language audit terminology, indexing documentary source excerpts, and verifying the scripts that render the comparative figures. All authors independently checked every source passage and visual output against the primary official reports, performed all substantive analysis and prose drafting themselves, and accept full responsibility for the published commentary.

Funding. Institutional research support was provided by the Academy of Public Policy and Economics (APPE). No external grant funding was received.

Competing interests. Samuel Tay, Nicholas Tang, and Sophie Hadfield are researchers at the Academy of Public Policy and Economics, publisher of Commentary and Critique. The authors declare no external competing financial or institutional interests.

References

Algemene Rekenkamer. (2023). *Resultaten verantwoordingsonderzoek 2022 Ministerie van Financiën en Nationale Schuld (IX)*. https://www.rekenkamer.nl/publicaties/rapporten/2023/05/17/resultaten-verantwoordingsonderzoek-2022-ministerie-van-financien

Algemene Rekenkamer. (2024). *Resultaten verantwoordingsonderzoek 2023 Ministerie van Financiën en Nationale Schuld (IX)*. https://www.rekenkamer.nl/publicaties/rapporten/2024/05/15/resultaten-verantwoordingsonderzoek-2023-ministerie-van-financien-en-nationale-schuld

Algemene Rekenkamer. (2025). *Resultaten verantwoordingsonderzoek 2024 Ministerie van Financiën en Nationale Schuld (IX)*. https://www.rekenkamer.nl/publicaties/rapporten/2025/05/21/resultaten-verantwoordingsonderzoek-2024-ministerie-van-financien-en-nationale-schuld

Alon-Barkat, S., & Busuioc, M. (2023). Human-AI interactions in public sector decision making: ”Automation bias” and ”selective adherence” to algorithmic advice. *Journal of Public Administration Research and Theory*, *33*(1), 153–169. https://doi.org/10.1093/jopart/muac007

Alon-Barkat, S., & Busuioc, M. (2024). Public administration meets artificial intelligence: Towards a meaningful behavioral research agenda on algorithmic decision-making in government. *Journal of Behavioral Public Administration*, *7*(1), 1–12. https://doi.org/10.30636/jbpa.71.261

Australian Government. (2023). *Government response to the Royal Commission into the Robodebt Scheme*. Department of the Prime Minister and Cabinet, Commonwealth of Australia. https://www.pmc.gov.au/resources/government-response-royal-commission-robodebt-scheme

Australian Public Service Commission. (2024). *Centralised Code of Conduct Inquiry Taskforce: Final report*. Commonwealth of Australia. https://www.apsc.gov.au/initiatives-and-programs/robodebt-code-conduct-process

Bovens, M., & Zouridis, S. (2002). From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review*, *62*(2), 174–184. https://doi.org/10.1111/0033-3352.00168

Bullock, J. B. (2019). Artificial intelligence, discretion, and bureaucracy. *The American Review of Public Administration*, *49*(7), 751–761. https://doi.org/10.1177/0275074019856123

Carney, T. (2019). Robo-debt illegality: The seven veils of failed guarantees of the rule of law? *Alternative Law Journal*, *44*(1), 4–10. https://doi.org/10.1177/1037969x18815913

Christensen, J., Aarøe, L., Baekgaard, M., & Herd, P. (2020). Human capital and administrative burden: The role of cognitive resources in citizen-state interactions. *Public Administration Review*, *80*(1), 127–136. https://doi.org/10.1111/puar.13134

Commonwealth Ombudsman. (2017). *Centrelink's automated debt raising and recovery system: A report about the Department of Human Services' Online Compliance Intervention system for debt raising and recovery (Report No. 02/2017)*. Office of the Commonwealth Ombudsman. https://www.ombudsman.gov.au/publications/investigation-reports/centrelinks-automated-debt-raising-and-recovery-system

Department for Work and Pensions. (2025). *Annual report and accounts 2024–25 (HC 995)*. His Majesty's Stationery Office. https://www.gov.uk/government/publications/dwp-annual-report-and-accounts-2024-to-2025

European Commission for Democracy through Law. (2021). *The Netherlands: Opinion on the legal protection of citizens (Opinion No. 1031/2021, CDL-AD(2021)031)*. Council of Europe. https://www.venice.coe.int/webforms/documents/?pdf=CDL-AD(2021)031-e

Henman, P. (2020). Improving public services using artificial intelligence: Possibilities, pitfalls, governance. *Asia Pacific Journal of Public Administration*, *42*(4), 209–221. https://doi.org/10.1080/23276665.2020.1816188

Herd, P., & Moynihan, D. P. (2019). *Administrative burden: Policymaking by other means*. Russell Sage Foundation. https://doi.org/10.7758/9781610448789

Meijer, A., Lorenz, L., & Wessels, M. (2021). Algorithmization of bureaucratic organizations: Using a practice lens to study how context shapes predictive policing systems. *Public Administration Review*, *81*(5), 837–846. https://doi.org/10.1111/puar.13391

Moynihan, D., Herd, P., & Harvey, H. (2015). Administrative burden: Learning, psychological, and compliance costs in citizen-state interactions. *Journal of Public Administration Research and Theory*, *25*(1), 43–69. https://doi.org/10.1093/jopart/muu009

National Audit Office. (2024). *Use of artificial intelligence in government (HC 612, Session 2023–24)*. https://www.nao.org.uk/reports/use-of-artificial-intelligence-in-government/

National Audit Office. (2025). *Tackling benefit overpayments due to fraud and error (HC 1336, Session 2024–26)*. https://www.nao.org.uk/reports/tackling-benefit-overpayments-due-to-fraud-and-error/

Parlementaire ondervraging kinderopvangtoeslag. (2020). *Ongekend onrecht: Verslag van de Parlementaire ondervraging kinderopvangtoeslag (Kamerstuk 35 510, nr. 2)*. Tweede Kamer der Staten-Generaal. https://www.tweedekamer.nl/kamerstukken/detail?id=2020D50966&did=2020D50966

Parliament of Australia. (1991). *Social Security Act 1991 (No. 46 of 1991, ss. 1064, 1223(1), and 1228B)*, Federal Register of Legislation. https://www.legislation.gov.au/C2004A04121/latest/text.

Parliament of Australia. (1999). *Social Security (Administration) Act 1999 (No. 191 of 1999, ss. 63, 80, 81, 192, 195, and 196)*, Federal Register of Legislation. https://www.legislation.gov.au/C2004A00580/latest/text.

Parliament of the United Kingdom. (2025). *Public Authorities (Fraud, Error and Recovery) Act 2025 (2025 c. 28)*, UK Public General Acts. https://www.legislation.gov.uk/ukpga/2025/28/enacted.

Peeters, R., & Widlak, A. C. (2023). Administrative exclusion in the infrastructure-level bureaucracy: The case of the Dutch daycare benefit scandal. *Public Administration Review*, *83*(4), 863–877. https://doi.org/10.1111/puar.13615

Raad van State. (2019). *Afdeling bestuursrechtspraak uitspraken van 23 oktober 2019 inzake kinderopvangtoeslag en evenredigheid (ECLI:NL:RVS:2019:3535 en ECLI:NL:RVS:2019:3536)*, Rechtspraak.nl. https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RVS:2019:3535.

Rechtbank Den Haag. (2020). *NJCM c.s. tegen de Staat der Nederlanden (SyRI-vonnis van 5 februari 2020, ECLI:NL:RBDHA:2020:1878)*, Rechtspraak.nl. https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2020:1878.

Redden, J., Dencik, L., & Warne, H. (2020). Datafied child welfare services: Unpacking politics, economics and power. *Policy Studies*, *41*(5), 507–526. https://doi.org/10.1080/01442872.2020.1724928

Royal Commission into the Robodebt Scheme. (2023). *Report of the Royal Commission into the Robodebt Scheme (Vols. 1–3)*. Commonwealth of Australia. https://robodebt.royalcommission.gov.au/publications/report

Staten-Generaal der Nederlanden. (2022). *Wet van 26 oktober 2022 houdende regels ten behoeve van de hersteloperatie toeslagen (Wet hersteloperatie toeslagen, Stb. 2022, 433)*, Staatsblad van het Koninkrijk der Nederlanden. https://wetten.overheid.nl/BWBR0047342.

Whiteford, P. (2021). Debt by design: The anatomy of a social policy fiasco – Or was it something worse? *Australian Journal of Public Administration*, *80*(2), 340–360. https://doi.org/10.1111/1467-8500.12479

Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, *12*(4), 505–523. https://doi.org/10.1111/rego.12158