The Authorisation Bottleneck

ENTERPRISE AI | DELEGATION | RENEWABLE AUTHORITY | QUEUEING CONTROL

ABSTRACT

Standing authority lets routine work proceed without repeated human approval. Its value depends on remaining usable: when the conditions of a mandate change, someone must review and renew it. If the same people also approve individual cases, clearing the immediate backlog can prevent the renewal that would reduce future review work. This paper models that conflict. A reviewer divides effort between cases and restoring a policy that authorises an eligible share of future arrivals. The model gives an exact capacity frontier and a strict lower bound on the protected renewal share within a constant-allocation policy class. Where renewal expands capacity, there is a range of workloads that protected renewal can sustain but strict case-first review cannot. Below ordinary review capacity, however, case-first renewal weakly improves backlog relative to ordinary review, while immediate renewal can increase delay. A finite backlog trigger retains the capacity guarantee while permitting case priority below the trigger; a worst-case certificate covers stated parameter uncertainty. Allocation must still reflect workload and service requirements. The results are proved in a minimal model and checked by separate numerical implementations. They generate testable predictions about renewal effort, policy availability and backlog; they do not establish organisational prevalence or decision quality.

KEYWORDS authorisation; delegation; policy renewal; queueing control; human review; enterprise AI.

1 The authorisation bottleneck and the claim

A standing policy can authorise many routine actions through one accountable decision. When its assumptions or mandate change, those actions return to individual review until the policy is renewed. If the people who renew policies also review cases, an organisation faces a choice that a visible backlog can obscure: process another case now, or restore the authority that would remove some future cases from the queue.

Consider routine software changes under a documented release policy. An environmental change suspends that policy. An existing manual process still permits individual changes to be reviewed, but automated release cannot resume until a reviewer revalidates the mandate. Clearing cases first appears responsive. At sufficiently high load, it can leave no time for revalidation and preserve the workload that prevents recovery. This example motivates the model; it is not evidence that any particular organisation satisfies its assumptions.

The claim is specific. When standing authority reduces enough future review work relative to the effort needed to restore it, protecting renewal time can sustain more activity than individual review alone. The same protection can be unnecessary for capacity, or costly in delay, at lower load. Authorisation is therefore both a control over work and, when it can be maintained economically, a source of processing capacity.

An organisational state account defines AI adoption through sustained delegation and the evidence that supports it [6]. Sustaining that delegation also requires usable authority. We examine this resource-allocation problem with a separate queueing model; we do not derive the adoption model's hysteresis from the queue.

This is an application of established authority and queueing ideas, not a claim that delegation or maintenance is new. Decision rights and management by exception already have a substantial literature [1, 3]; constrained human approval, action-dependent service and workload-dependent maintenance also have close precedents [11, 14, 15]. The contribution is a controlled model in which the reviewer restores a routing state, together with explicit capacity, backlog-trigger and uncertainty results. Appendix A defines that contribution boundary; Appendices B–G contain the model, proofs, numerical checks, empirical design, duration extension and reproducibility details.

2 Standing authority as a maintained resource

Fix a class of cases and a standing mandate under which an eligible case can proceed without individual review. The policy is either valid or suspended. Invalidation is observed when it occurs. While valid, the mandate authorises an independently eligible share of new cases immediately; the remainder enter the manual queue. While suspended, all arrivals enter that queue. Renewal restores the policy for future arrivals. It does not retrospectively release cases already waiting, and no case is discarded.

One reviewer supplies the effort for both case review and renewal. Candidate arrivals have rate λ; full-effort case review and renewal have rates μ and ν; valid policies are suspended at rate δ; and the eligible share is x. The baseline model

assumes Poisson arrivals and exponential effort requirements and valid lifetimes. It isolates the allocation conflict by holding these quantities fixed within each comparison.

Validity means permission to use the mandate, not proof that its decisions are correct. Both the standing policy and the manual fallback must be independently admissible. Hidden degradation, misclassified eligibility and harmful approvals require additional models of detection and loss. Better technology may change case productivity, renewal productivity, eligible scope or policy lifetime. The comparison concerns the resulting system, not an assumption that human review productivity can never improve.

3 The renewal and backlog mechanism

The mechanism has two directions. A valid policy removes eligible arrivals from manual review. Suspension restores those arrivals, while renewal competes with the cases now waiting. The question is whether the reviewer can restore authority often enough to offset its maintenance cost.

The model's supremum stabilisable candidate-arrival rate is

$$\lambda_{\text{cap}} = \max\left\{\mu, \frac{\mu}{1 - x + \delta/\nu}\right\}. \tag{1}$$

Renewal expands capacity exactly when $\nu x > \delta$. The condition compares eligible scope with renewal effort per unit of valid-policy lifetime. It includes the bypass opportunities lost while the policy is suspended. A stationary resource bound limits what any admissible allocation with stationary occupation measures can achieve; a separate stability proof constructs an allocation that achieves every strict load below the frontier (Theorem C.1).

For a policy that reserves a constant share h of reviewer effort whenever the mandate is suspended, the exact stability condition yields a critical reservation share $h_{\min}$ at loads above μ (Theorem C.2). The reservation must exceed that threshold. Reserving time only after every case has been processed does not suffice: above μ, an invalid policy with a non-empty queue has a positive probability of never reaching an empty queue. Renewal then never resumes. Thus there is a non-empty interval of loads that protected renewal stabilises and strict case-first review does not (Theorem C.3).

A constant reservation is not necessary at every positive backlog. A rule that renews when idle, prioritises cases below a finite queue threshold and protects renewal above it retains the same sufficient capacity guarantee (Theorem C.5). A separate worst-case certificate identifies sufficient protection when the fixed parameters are known only to lie in stated ranges (Proposition C.1).

The reverse comparison matters at low load. Below μ, case-first renewal uses only otherwise idle effort and never gives waiting cases less service than ordinary review. Its backlog is stochastically no larger than the baseline backlog (Theorem C.4). Immediate renewal can instead interrupt useful case service. Capacity and responsiveness need not favour the same allocation.

At the constructed values $\mu = 3$, $\nu = 1$, $\delta = 0.4$ and $x = 0.6$, the capacity frontier rises from 3 to 3.75 cases per week. At $\lambda = 3.4$, the invalid-time reservation must exceed $4/11$. A half reservation is stable but produces mean delay of about 24.75 weeks; immediate renewal reduces that to 8 weeks. At moderate load, a backlog trigger can improve on immediate renewal; at high load it can increase delay (Appendix D.6). Stability is therefore a feasibility condition, not an acceptable service target. Figure 1 summarises the mechanism and reservation boundary; Appendix D reports the checks and uncertainty.

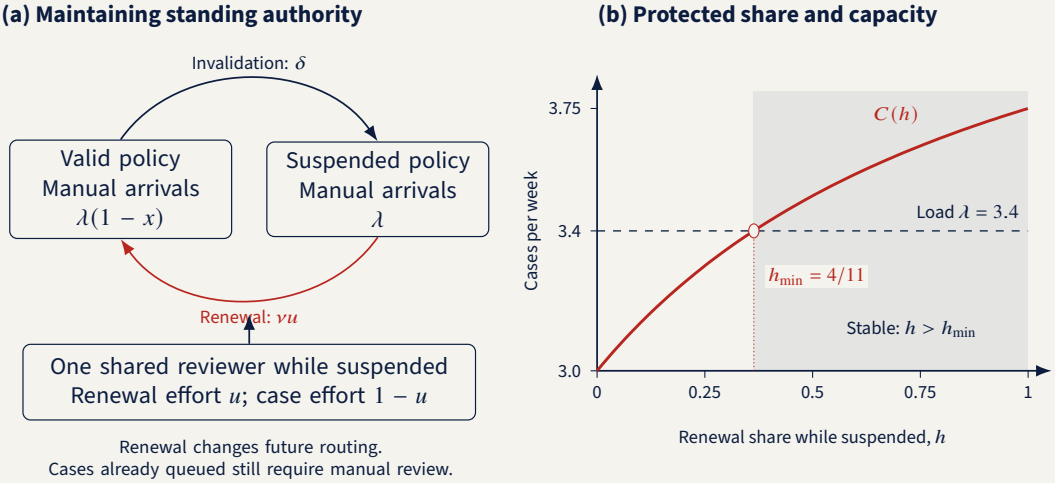

Figure 1. Renewable authority and protected capacity. The left panel shows the shared-effort mechanism. The right panel uses the constructed parameters in the text: at $\lambda = 3.4$, fixed reservation is stable only to the right of $h_{\min} = 4/11$. These are model quantities, not estimates from an organisation.

4 Empirical implications and failure conditions

The distinguishing prediction is not simply that delegation reduces approvals. It is that renewal allocation explains whether authority remains available as workload rises. Above measured ordinary review capacity, case-first renewal should become unreliable; a sufficiently protected share should preserve policy availability and prevent sustained backlog growth. Below that capacity, the delay comparison can favour renewal performed in idle time.

A prospective study should measure candidate arrivals, eligible scope, policy invalidation, active renewal effort, case-review effort and queue occupancy. Calendar time to renewal includes waiting and must not be treated as active renewal effort. Estimate the parameters in one period and test the predicted load range and reservation boundary in a later period. Compare assisted case review, case-first renewal and protected renewal with the same total reviewer budget. Assignment should occur at the shared-resource level where feasible, because cases sharing a reviewer interfere with one another.

The institutional explanation is weakened if policy suspension does not materially increase manual arrivals, if renewal is not competing for the same resource, or if protected allocation fails to improve availability and backlog where independently estimated parameters predict it should. Such findings would challenge the model's application, even if its conditional mathematics remains valid. If invalidation cannot be observed, or manual fallback is not admissible, the proposed operating comparison does not apply.

Faster authorisation must also preserve the required quality of decisions. Measure harmful approvals, missed escalation, valid rejection and useful completion, including unfinished cases and full maintenance costs. A short observation without backlog growth does not prove stability; absence of a significant quality difference does not establish equivalence. Appendix E sets out the study design and institutional limits. No field experiment or organisational treatment effect is claimed here.

5 Conclusion

Standing authority only reduces review work while it remains usable. When restoring it draws on the same people who process cases, protecting renewal capacity can sustain workloads that strict case priority cannot. A finite backlog trigger can preserve case priority at small queues without relying on spare time indefinitely. The parameter-uncertainty certificate specifies when that capacity claim survives a stated range of estimates. The next empirical task is to test that allocation mechanism in workflow and policy records, with decision quality and service requirements assessed alongside throughput.

A Evidence base and contribution boundary

Aghion and Tirole distinguish formal authority from effective control and analyse information and overload in authority allocation [1]. Athey and colleagues study state-contingent decision rights and management by exception [3]. These precedents rule out novelty claims for distinguishing entitlement from attention or escalating exceptions while delegating routine decisions.

Research on human–AI service operations already models constrained human review. Lin and Zhang analyse AI production, screening and mandatory human release approval with rework [15]. Lykouris and Weng jointly consider classification, admission, scheduling and learning from delayed human feedback [17]. Our focus is different: an eligible bypass remains available only while a policy is valid, and the reviewer must restore that policy after invalidation. Eligibility and decision quality are inputs here, not learned optimal classifications.

Maintenance and controlled service systems are equally important precedents. Lin, La and Martins characterise stabilisability when service performance depends on activity and rest [14]. Their work already establishes that serving work whenever possible need not maximise sustainable throughput. Kaufman and Lewis also analyse workload-dependent maintenance, including settings in which repair is favoured at an empty queue and again at larger backlogs [11]. Low-load idleness and high-load maintenance are therefore not new qualitative observations. Our distinguishing state controls arrival routing rather than the intrinsic productivity of case service. Federgruen and So prove an optimal repair threshold for a broken-server model with holding, repair and running costs [9]. Their server cannot process cases while waiting for repair; our manual fallback remains available while the routing policy is invalid. We establish stability for specified allocations, not their globally optimal cost. The general principle of investing in future performance and the use of congestion thresholds are established.

State-dependent arrivals during breakdowns are also established [18]. Working-vacation models allow different arrival and service rates across operating states [10], and Markov-modulated service control treats the interaction between congestion and the environment [13]. Once a fixed renewal allocation is chosen, our queue belongs to this family. Its stationary mean is therefore a specialised calculation within an established queueing class. The control question is whether to restore the routing state, how much shared effort to reserve, and when case-first renewal prevents capacity expansion. No priority claim is made for the fixed-policy mean against the breakdown-queue literature.

Koyanagi and Kawai study maintenance when demand depends on a deteriorating server state [12]. Here external candidate demand remains fixed, all cases are retained, and renewal reduces the share requiring manual service. This comparison matters: attributing novelty to maintenance changing arrivals would be too broad.

Software-delivery research already recommends combining peer review, automation and explicit approval-wait measurement [7]. The empirical contribution still to be made is to test whether measured policy invalidation and shared renewal effort explain a failure that simpler review assistance does not resolve. The present paper supplies model results and an evaluation design; it does not establish that this mechanism is prevalent in organisations.

B Model and admissible controls

B.1 States and transitions

Candidate cases arrive as a Poisson process of rate $\lambda > 0$. A single reviewer supplies one unit of effort per unit of calendar time. A manually reviewed case requires exponentially distributed effort with mean $1/\mu$, where $\mu > 0$. All case requirements, renewal requirements, validity lifetimes and eligibility marks are mutually independent and independent of the candidate-arrival process. Requests are processed first in, first out, with preemptive resume when renewal receives effort. Counts include cases in service.

Let $Q \in \{0, 1, 2, \ldots\}$ be the number of manual cases waiting or in service. Let $V = 1$ mean that a specified standing policy is currently valid, and $V = 0$ mean that its use is suspended. Independently of case arrivals and service, a valid policy becomes invalid at rate $\delta > 0$. Invalidation is observed immediately. While the policy is valid, each arriving case independently belongs to the policy's eligible class with probability $x \in [0, 1]$. Eligible cases receive immediate authorisation; the remaining cases require manual review. While the policy is invalid, every arriving case requires manual review.

No case is discarded. Cases already in the manual queue remain there when renewal completes; the model does not retrospectively release them. Consequently the manual-arrival rate is $\lambda(1 - xV)$. The external candidate-arrival rate remains λ under every policy. Automated execution capacity and post-authorisation execution time are outside the model.

Renewal requires exponentially distributed reviewer effort with mean $1/\nu$, where $\nu > 0$. If fraction $u \in [0, 1]$ of effort is allocated to renewal while $V = 0$, renewal completes at rate νu and available case service has rate $\mu(1 - u)$. Renewal cannot occur while $V = 1$. Case review then has full rate μ when the queue is non-empty. Fractional effort is an idealisation of arbitrarily fine time sharing; it does not assert that a person performs two tasks simultaneously.

Event	Condition and rate	Effect
Manual arrival	λ when invalid; $\lambda(1-x)$ when valid	Q increases by one
Case completion	$\mu(1-u)$ when invalid; μ when valid, provided $Q > 0$	Q decreases by one
Policy invalidation	δ when $V = 1$	V becomes 0
Renewal completion	νu when $V = 0$	V becomes 1

Table 1. Queue transitions under shared review capacity.

B.2 What the validity state means

Validity means that the model permits eligible cases to use the standing mandate. It does not imply that the mandate is substantively correct, lawful in every setting or safe. We assume that the eligible class and fallback process have already been found acceptable under the organisation's governing requirements. The model compares review-resource consequences within that admissible set.

Immediate observation of invalidation is material. If an organisation cannot reliably detect when a mandate ceases to apply, the model cannot establish the safety of continued use. A policy with hidden degradation requires a model of detection and decision loss. Similarly, state-dependent case difficulty, imperfect eligibility classification, renewal failure, rework and interference between several policies are not silently absorbed into the proofs.

The rates μ and ν describe the same reviewer at full allocated effort under a specified assistance configuration. Changing technology may change either rate and may change x or δ. A comparison that holds all four fixed while claiming to estimate the effect of AI would be mis-specified. Staffing, monitoring and supervision commitments must be reflected in measured available capacity; setup costs are evaluated separately from long-run stability.

B.3 Controls and the meaning of capacity

An admissible control observes the current queue and validity state, cannot foresee random future events, cannot use an invalid policy and cannot delete cases or create additional review capacity. It may allocate effort between cases and renewal while invalid. We allow the baseline option of declining renewal and using ordinary review permanently. The upper bound concerns stabilising regimes with stationary occupation measures, including randomised state-dependent controls. The constructive policies below are stationary.

For a stationary Markov control, we call a regime stable when its accessible recurrent operating class is positive recurrent, so the queue and validity state have an invariant probability distribution. For the broader upper bound, stability means a proper stationary regime with stationary state-and-action occupation measures. We do not assert a converse for arbitrary non-stationary notions of stability. The case-only baseline has an eventually invalid policy and a stable M/M/1 queue when $\lambda < \mu$. Our capacity frontier is the supremum of stabilisable λ. We prove achievability strictly below the frontier and impossibility strictly above it; no operating recommendation relies on equality at the boundary.

The aim is sustainable processing of all arriving cases, not profit maximisation or a globally optimal delay policy. This distinction is necessary because a rule can maximise the capacity frontier while imposing unnecessary waits at low load.

C Capacity allocation and proofs

C.1 The maximal stabilisable arrival rate

Theorem C.1. *Under Appendix B's assumptions, the supremum stabilisable candidate-arrival rate is:*

$$\lambda_{\text{cap}} = \max\left\{\mu, \frac{\mu}{1 - x + \delta/\nu}\right\}. \tag{C.1}$$

If $\nu x \le \delta$, renewal cannot expand the capacity frontier beyond ordinary review. If $\nu x > \delta$, immediate renewal while invalid attains stability for every λ below $\mu/[1 - x + \delta/\nu]$. This second policy gives renewal full priority while invalid and reviews cases while valid.

The bound accounts for lost bypass opportunities during renewal as well as the renewal effort itself. The expansion condition compares eligible share x with renewal effort per unit of valid-policy lifetime, δ/ν. Improving case productivity μ scales the frontier; improving renewal productivity ν or extending the validity period can also determine whether a capacity expansion is possible at all.

The proof is given in Appendix C.5. Its key quantities are the stationary valid fraction v and the stationary renewal-effort fraction r. Validity balance gives $r = \delta v/\nu$, and renewal can use only invalid time, so $v \le \nu/(\nu+\delta)$. Case-flow balance and the unit effort budget imply:

$$\frac{\lambda(1-xv)}{\mu} + \frac{\delta v}{\nu} \le 1. \tag{C.2}$$

Maximising the resulting bound over feasible v gives equation (C.1). This occupation bound alone would not establish achievability. A separate drift construction proves stability of the policy that reaches the favourable endpoint.

C.2 How much time must be protected

For a constant reservation $h \in (0,1]$, allocate fraction h to renewal whenever invalid and retain fraction $1-h$ for cases. While valid, give cases full capacity. Reserved renewal effort remains fixed even when the manual queue is empty; unused case capacity is not reassigned. This restriction makes the policy transparent and its formula exact.

Theorem C.2. *A constant reservation $h \in (0,1]$ is stable if and only if:*

$$\lambda < C(h) = \frac{\mu[\delta + h(\nu-\delta)]}{\delta + h\nu(1-x)}. \tag{C.3}$$

Its capacity is increasing in h exactly when $\nu x > \delta$, decreasing when $\nu x < \delta$, and constant when the two are equal. For $\mu < \lambda < \lambda_{\text{cap}}$ in the expansion regime, the exact strict reservation requirement within this policy class is:

$$h > h_{\min} = \frac{\delta(\lambda-\mu)}{\mu(\nu-\delta) - \lambda\nu(1-x)}. \tag{C.4}$$

Here $0 < h_{\min} < 1$. The threshold is an infimum, not an attained stable minimum: equality does not suffice. The threshold concerns the fraction reserved while invalid, not the fraction of all reviewer time spent renewing. The latter is $h\delta/(\delta+\nu h)$. A reservation can stabilise the queue and still leave poor delay performance if it lies close to $h_{\min}$.

The formula can reject a proposed allocation before deployment. It does not justify choosing h at equality, rounding down an uncertain estimate or ignoring other work. When parameters are estimated, require a positive slack margin across a joint plausible parameter set. The margin and the set must be justified from data rather than chosen to reproduce a desired decision.

C.3 Why strict case priority can fail

Strict case-first review serves cases whenever $Q > 0$ and works on renewal only when the policy is invalid and $Q = 0$. Renewal is immediately preempted by a new case. This is an appealing response to a visible backlog.

Theorem C.3. *Strict case-first review is not positive recurrent when $\lambda > \mu$. Starting from an invalid policy with $Q = n \ge 1$, its probability of never reaching an empty queue is $1 - (\mu/\lambda)^n$. On that event renewal never resumes. Therefore, when $\nu x > \delta$, every λ in $(\mu, \lambda_{\text{cap}})$ is a load for which protected renewal can stabilise the system but strict case-first review cannot.*

The failure does not require renewal to be useless or intrinsically slow. Once invalid, a non-empty queue under strict case priority is an ordinary overloaded M/M/1 queue. Its positive probability of never emptying prevents the very intervention that would lower future manual arrivals. The proof uses the standard birth–death hitting probability, not a simulated backlog explosion.

C.4 The same rule can help below ordinary capacity

Theorem C.4. *For $\lambda < \mu$, strict case-first renewal has a queue-length process stochastically no larger than ordinary case-only review, under a coupling with common candidate arrivals and common potential case-service completions, starting with equal queues. Its stationary mean authorisation delay averaged across all candidate cases is therefore no larger than $1/(\mu-\lambda)$.*

When there are cases to serve, strict case priority always gives them the same full service capacity as ordinary review. A valid policy can remove some new manual arrivals and never adds them beyond the ordinary baseline. Renewal consumes only otherwise idle effort. This establishes the coupling; it does not claim that strict case priority minimises delay among all controls.

Theorems C.3 and C.4 explain why a successful low-load practice can become a poor scaling rule. Below μ it can exploit idle time without delaying case service. Above μ, waiting for idle time can prevent renewal indefinitely. A decision to reserve capacity therefore depends on load and objective, not only on whether delegation has worked before.

C.5 Occupation bound and achievability

Let $v = P(V = 1)$, r be average effort devoted to renewal and c be average effort actually processing cases in a stationary regime. Stationary validity balance gives $\delta v = \nu r$. Since renewal occurs only while invalid, $r \le 1 - v$. Stationary case flow gives $\mu c = \lambda(1 - xv)$; bounded transition intensities permit this balance by summing level-crossing identities, without assuming a finite mean queue in advance. The effort constraint $c + r \le 1$ gives equation (C.2).

Thus $\lambda \le F(v) = \mu(1 - \delta v/\nu)/(1 - xv)$, with $0 \le v \le \nu/(\nu + \delta)$. The denominator is strictly positive because $\delta > 0$. Its derivative has the sign of $x - \delta/\nu$. If that sign is non-positive, F is maximised at $v = 0$ and equals μ. Otherwise its maximum is at $v = \nu/(\nu + \delta)$ and equals $\mu/[1 - x + \delta/\nu]$. This establishes an upper bound for every admissible stationary regime, not only for the constant-reservation family.

For the constant-reservation family $h \in (0, 1]$, the environment changes from invalid to valid at rate νh and back at δ, with rates that do not depend on Q. For $Q > 0$ use $f(Q, V) = Q + b(1 - V)$, with a constant b chosen positive. The two generator drifts are $\lambda - \mu(1 - h) - \nu h b$ when invalid, and $\lambda(1 - x) - \mu + \delta b$ when valid. Both are strictly negative if:

$$\frac{\lambda - \mu(1 - h)}{\nu h} < b < \frac{\mu - \lambda(1 - x)}{\delta}. \tag{C.5}$$

Under $\lambda < C(h)$, the right endpoint is positive and the interval is non-empty, so such a positive b exists. Transition rates are bounded and the chain is irreducible on its recurrent class. The Foster drift condition outside the finite set $Q = 0$ gives positive recurrence [2]. Applying the same argument to $\exp(\theta f)$ for sufficiently small $\theta > 0$ gives finite queue moments, because the jumps are bounded and both linear drifts are strictly negative. This also justifies the stationary moment calculations below.

At $h = 1$, $C(h) = \mu/[1 - x + \delta/\nu]$. The case-only option stabilises every $\lambda < \mu$. Together these constructions attain every strict load below equation (C.1), proving that the upper bound is the supremum. They do not assert finite delay at equality.

C.6 The reservation threshold

The invariant valid fraction for a fixed $h > 0$ is $\nu h/(\delta + \nu h)$. The averaged potential case-service rate is $\mu[1 - h\delta/(\delta + \nu h)]$, and the averaged manual-arrival rate is $\lambda[1 - x\nu h/(\delta + \nu h)]$. Their strict inequality simplifies to $\lambda < C(h)$, matching Appendix C.5's sufficient drift condition. The corresponding non-strict inequality is necessary by stationary flow balance. At equality, this irreducible Markov queue cannot have a stationary law: the balance would require zero unused potential case service, but the valid empty state must have positive stationary probability. Above equality the necessary balance fails. Thus the strict condition is exact for this family.

Differentiating $C(h)$ gives a numerator $\mu\delta(\nu x - \delta)$ over the positive squared denominator. For $\mu < \lambda < \lambda_{\text{cap}}$ in the expansion regime, rearrangement yields $h > h_{\min}$ in equation (C.4). The denominator is positive because at λ_{cap} it equals $\delta(\lambda_{\text{cap}} - \mu) > 0$ and it is non-increasing in λ. The same comparison gives $h_{\min} < 1$; positivity is immediate from $\lambda > \mu$. At $h = 0$, eventual invalidity gives the case-only baseline, treated separately from Appendix C.5.

C.7 Case-first starvation and low-load domination

When invalid with $Q > 0$, strict case-first service uses all effort on cases. Until Q reaches zero, its birth and death rates are λ and μ and renewal rate is zero. For $\lambda > \mu$, the probability of hitting zero from n is $(\mu/\lambda)^n$, obtained by solving the birth–death hitting-probability recurrence with the minimal non-negative solution [2]. There is positive probability of escape without renewal. The accessible operating class contains every invalid state and communicates with $(0, 1)$, so it cannot be positive recurrent. When $x = 1$, valid states with $Q > 0$ are transient and need not belong to that class; this does not affect the escape argument. This proves Theorem C.3. No conclusion at $\lambda = \mu$ is needed for that theorem.

For $\lambda < \mu$, couple a case-only queue and a case-first queue using one candidate Poisson process and one potential case-completion Poisson process of rate μ. Every candidate enters the baseline; only the candidates requiring manual review enter the case-first queue. Whenever that queue is non-empty, renewal is preempted and every potential case completion is usable. Induction over event times gives $Q_{\text{case-first}} \le Q_{\text{baseline}}$ from equal initial counts. Policy-state transitions cannot increase either queue. The same function $f(Q) = Q$ has drift at most $\lambda - \mu < 0$ outside $Q = 0$, establishing positive recurrence of the case-first chain; exponential drift gives finite mean. The stationary order follows by the coupling, and Little's law with external rate λ gives Theorem C.4. The comparison excludes renewal setup and monitoring costs not already represented in capacity.

C.8 Protecting renewal only above a backlog threshold

Permanent reservation is sufficient but is not the only way to prevent starvation. Fix a finite integer $K \ge 1$ and a share $h \in (0, 1]$. While the policy is invalid, renew at full effort when the queue is empty, give cases priority when $1 \le Q < K$, and reserve share h when $Q \ge K$. Thus

$$u_K(Q) = \begin{cases} 1, & Q = 0, \\ 0, & 1 \le Q < K, \\ h, & Q \ge K. \end{cases} \tag{C.6}$$

Here u_K denotes allocated effort, not the renewal rate νu_K. The allocation is re-evaluated after each event. In particular, if case completion takes the queue below K while $h < 1$, renewal is preempted. Service remains first in, first out with preemptive resume. The rule requires no prediction of future arrivals.

Theorem C.5. *For every finite $K \geq 1$, the backlog-triggered policy in (C.6) is positive recurrent and has a finite exponential queue moment whenever $\lambda < C(h)$. In the capacity-expansion regime, setting $h = 1$ therefore attains every strict load below the frontier (C.1), for any fixed finite K.*

Proof. Choose $b > 0$ satisfying (C.5). On every state with $Q \geq K$, the generator of $f(Q, V) = Q + b(1 - V)$ is exactly that of the constant-reservation policy. Both drifts are strictly negative. All changed allocations lie in the finite set $Q < K$, where transition rates and generator increments are bounded. The chain is irreducible: positive arrivals can reach the renewal region; renewal permits valid-state case completions; and invalidation reconnects the two policy states. The same finite-set Foster argument as in Appendix C.5 gives positive recurrence. Bounded jumps make the generator of $\exp(\theta f)$ negative outside that finite set for sufficiently small $\theta > 0$, establishing a finite exponential moment. For $h = 1$ in the expansion regime, $C(1) = \lambda_{\text{cap}}$; the universal upper bound is unchanged. ∎

This result uses an established finite-set drift argument [2]. Threshold repair and the pattern of repairing at an empty queue and again at high congestion have close maintenance precedents [9, 11]; neither pattern is claimed as a new general principle. The result establishes their capacity guarantee in this authority-routing model. It removes the need to protect a constant share at every positive backlog. It does not prove that a particular K minimises delay or cost. A large finite threshold can impose long waits before renewal, even though it preserves the asymptotic guarantee. Strict case-first renewal corresponds to never entering a high-backlog renewal region; Theorem C.3 shows why the finite-threshold result cannot simply be extended to $K = \infty$ above ordinary capacity.

C.9 A certificate under parameter uncertainty

A point estimate close to the capacity boundary is not an allocation guarantee. Suppose the true, fixed parameter vector lies in a stated rectangular set with positive rate bounds,

$$\lambda \in [\lambda_-, \lambda_+], \quad \mu \in [\mu_-, \mu_+], \quad \nu \in [\nu_-, \nu_+], \quad \delta \in [\delta_-, \delta_+], \quad x \in [x_-, x_+] \subseteq [0, 1]. \tag{C.7}$$

The set must be justified independently of the desired allocation. It may be a conservative enclosure of a joint uncertainty region; it is not automatically a confidence region merely because it is written as intervals.

Proposition C.1. *For a fixed $h \in (0, 1]$, define*

$$C_-(h) = \frac{\mu_-[\delta_+ + h(\nu_- - \delta_+)]}{\delta_+ + h\nu_-(1 - x_-)}. \tag{C.8}$$

If $\lambda_+ < C_-(h)$, the constant-reservation policy and every fixed finite-threshold policy (C.6) are stable for each fixed parameter vector in (C.7). If $\mu_- < \lambda_+ < C_-(1)$, a sufficient strict reservation requirement is

$$h > h_{\text{rob}} = \frac{\delta_+(\lambda_+ - \mu_-)}{\mu_-(\nu_- - \delta_+) - \lambda_+\nu_-(1 - x_-)}. \tag{C.9}$$

For constant reservation, $\lambda_+ < C_-(h)$ is also necessary to guarantee stability at every point of the closed rectangular set, which includes its worst corner.

Proof. The capacity $C(h)$ increases with μ, ν, x and decreases with δ for fixed $h > 0$, with possible equality in degenerate cases. Direct differentiation of (C.3) gives these signs. Its minimum on the rectangle is therefore $C_-(h)$. Theorem C.2 proves the constant-policy claim, including necessity at the attained worst corner. For the threshold policy, choose a single positive b strictly between

$$\frac{\lambda_+ - \mu_-(1 - h)}{\nu_- h} \quad \text{and} \quad \frac{\mu_- - \lambda_+(1 - x_-)}{\delta_+}.$$

Such a b exists under $\lambda_+ < C_-(h)$. For every parameter vector in the set, the invalid-state drift is at most $\lambda_+ - \mu_-(1-h) - \nu_- hb < 0$ and the valid-state drift is at most $\lambda_+(1 - x_-) - \mu_- + \delta_+ b < 0$. Apply Theorem C.5's finite-set argument. Rearranging the worst-corner inequality under the stated expansion conditions gives (C.9). ∎

The statement concerns uncertainty about fixed Markov parameters. It does not establish validity under arbitrary regime changes, hidden invalidation, or a data-dependent threshold that grows without bound. Nor does it certify a service-level target. As a constructed sensitivity example, take $\lambda \in [3.2, 3.4]$, $\mu \in [2.94, 3.06]$, $\nu \in [0.98, 1.02]$, $\delta \in [0.392, 0.408]$ and $x \in [0.588, 0.612]$. Then $h_{\text{rob}} \simeq 0.6076$, compared with the nominal threshold $4/11$. A share of 0.75 gives worst-corner capacity about 3.4619, exceeding the maximum load by about 0.0619 cases per week. A nominal half reservation cannot certify the whole box. These intervals are illustrative, not estimated organisational uncertainty.

D Delay and computational verification

D.1 Delay comparisons

Under immediate renewal, let $A = \nu(1 - x) + \delta$. The valid fraction is $v = \nu/(\nu + \delta)$. For $\lambda < \nu\mu/A$, the mean authorisation delay across all candidate cases is:

$$W_R = \frac{A + \delta(\mu + \lambda x)/(\nu + \delta)}{\nu\mu - \lambda A}. \tag{D.1}$$

Little's law divides mean queue length by λ, the external candidate rate: eligible cases authorised immediately contribute zero time. The mean for manually reviewed cases alone instead divides the stationary queue population by $\lambda(1 - xv)$. These two reporting populations must not be mixed. Equation (D.1) follows from stationary first and second queue moments; Appendix D.2 gives the derivation. It is a fixed-policy calculation in an established modulated queue family [2, 10, 18].

For $\lambda < \mu$ when immediate renewal is also stable, comparing equation (D.1) with ordinary review yields:

$$W_R < \frac{1}{\mu - \lambda} \quad \Longleftrightarrow \quad \mu(\nu x - \delta) > \frac{\delta(\mu + \lambda x)(\mu - \lambda)}{\nu + \delta}. \tag{D.2}$$

Even when $\nu x > \delta$, this inequality can fail at low load. The right-hand side decreases with λ on $(0, \mu)$, since its derivative has the sign of $\mu(x - 1) - 2x\lambda$. There is consequently at most one crossing from ordinary review being faster to immediate renewal being faster. This comparison is between two named policies; it is not a globally optimal switching threshold.

As a constructed example, take $\mu = 3$, $\nu = 1$, $\delta = 0.4$ and $x = 0.6$, with time measured in weeks. Ordinary review can sustain rates below 3; the frontier with renewal is 3.75. At $\lambda = 0.5$, ordinary mean delay is 0.4000 weeks while immediate renewal gives 0.6703. Strict case-first renewal gives 0.2289 by numerical stationary calculation. At $\lambda = 2.8$, the corresponding means are 5.0000, 2.8120 and 4.4482 weeks. The ranking changes with load.

At $\lambda = 3.4$, ordinary review and strict case-first renewal are unstable. Equation (C.4) requires $h > 4/11$, approximately 0.3636, while the policy is invalid. A reservation of 0.5 gives capacity 3.5 and mean delay approximately 24.7539 weeks; immediate renewal gives capacity 3.75 and mean delay 8.0000 weeks. Both are stable, yet neither delay should be called acceptable without a decision-specific requirement. This example exposes the difference between crossing a mathematical stability boundary and meeting an operational service target.

D.2 Derivation of the immediate-renewal mean

Write $M = E[Q]$, $M_1 = E[QV]$, $p = \nu/(\nu + \delta)$, and $\bar{\alpha} = \lambda(1 - xp)$. Case-flow balance gives $\mu P(Q > 0, V = 1) = \bar{\alpha}$. The stationary generator identities for Q^2 and QV yield respectively:

$$\begin{aligned} (\mu + \lambda x)M_1 &= \lambda M + \bar{\alpha}, \\ (\nu + \delta)M_1 &= \nu M - \lambda(1 - p). \end{aligned} \tag{D.3}$$

Substitution gives $M = \lambda[A + \delta(\mu + \lambda x)/(\nu + \delta)]/[\nu\mu - \lambda A]$. Since immediate authorisations remain in the external arrival cohort with zero queue time, $W = M/\lambda$. This establishes equation (D.1). Finite moments follow from Appendix C.5. Cross-multiplication by the positive denominators gives equation (D.2). The derivative stated in Appendix D.1 establishes at most one crossing between these two policies.

D.3 Separate stationary calculations

The proofs establish the model results. Computation checks their implementation and illustrates their consequences; it does not validate an organisation.

A Python implementation constructs a finite continuous-time generator directly from the transition table and solves its stationary balance equations. Queue cutoffs are increased and compared. A separate two-phase matrix-geometric calculation handles the unbounded fixed-policy queue. These methods agree with equation (D.1) for immediate renewal and with one another for the other stable policies. The largest displayed case, $h = 0.5$ at $\lambda = 3.4$, uses a finite cutoff of 1,600: its mean differs from the matrix solution by about 0.0000041 weeks, with probability about 0.000000000195 at that cutoff. Small cutoffs materially bias this example.

Two thousand independently sampled parameter sets check the occupation-bound maximisation, endpoint capacities and reservation identity. The domain is μ, ν, δ in $[\exp(-2), \exp(2)]$ and x in $[0, 1]$, sampled as specified in the supplied code. These sampled checks supplement the proofs; they are not evidence of universal validity by themselves.

D.4 Separate event simulation

A separate C++ implementation simulates competing continuous-time events without evaluating the analytic mean or solving stationary equations. Ten prespecified scenarios each use 20 independent runs. Every run discards the first 10,000 time units and measures the next 100,000. Seeds are recorded. The estimator is the time integral of Q divided by the measured duration and external rate λ. This avoids reporting only the cases that happen to complete before a simulation ends.

Policy and candidate rate	Stationary mean from analysis	Simulation mean and approximate 95 per cent interval
Immediate renewal, $\lambda = 0.5$	0.6703	0.6710 [0.6651, 0.6769]
Immediate renewal, $\lambda = 2.8$	2.8120	2.7741 [2.7333, 2.8150]
Immediate renewal, $\lambda = 3.4$	8.0000	7.8823 [7.6028, 8.1618]
Half reservation, $\lambda = 3.4$	24.7539	23.0509 [21.0327, 25.0692]
Strict case first, $\lambda = 0.5$	0.2289	0.2292 [0.2281, 0.2302]
Strict case first, $\lambda = 2.8$	4.4482	4.4601 [4.3444, 4.5759]

Table 2. Stationary means and event-simulation checks. Times are weeks for the constructed parameters.

All displayed times are weeks for the constructed parameters. Intervals use a Student t multiplier with 19 degrees of freedom across independent run means. They are pointwise approximate Monte Carlo intervals, not simultaneous guarantees and not bounds on warm-up bias, parameter uncertainty or organisational model error. The relatively wide half-reservation interval is retained rather than hidden: proximity to capacity makes estimation difficult.

The remaining scenarios are unstable by the analytic conditions. At $\lambda = 3.4$, strict case-first review has average measured backlog growth of 0.3961 cases per week, compared with asymptotic growth $\lambda - \mu = 0.4$ after renewal becomes starved. Reserving only one quarter produces measured growth of 0.0765, while its capacity is 3.3. Immediate renewal at $\lambda = 3.8$ and a high-invalidation case with $\delta = 1$ and $\lambda = 2.8$ (all other parameters unchanged) are also included. Their finite-run queue integrals are not reported as stationary delays. Observing growth is consistent with the proofs but does not itself prove instability.

D.5 Duration sensitivity

An additional 120 independent renewal-cycle runs each generate 100,000 cycles. Within each scenario, valid and renewal durations are independent and use the same distribution family: deterministic, exponential or gamma with shape 0.25, with means held at 2.5 and 1 week. Independently generated case requirements have mean 1/3 week. Equation (F.1) gives expected net work per cycle of -0.2333 at $\lambda = 3.4$ and $+0.0333$ at $\lambda = 3.8$. Across the three duration laws, the observed averages over runs range from -0.2349 to -0.2256 and from 0.0311 to 0.0355, respectively. All six aggregate estimates retain the predicted drift sign. Distributional variation changes sampling dispersion without changing the theoretical mean drift; the run-level estimates are retained as generated.

These checks concern cycle work, not a comparison of mean waiting times across those duration laws. The general-duration proposition supplies the stability implication. Its finite-mean limitations remain important for interpretation.

D.6 Backlog-triggered renewal: verification and trade-offs

For the constructed parameters $\mu = 3$, $\nu = 1$, $\delta = 0.4$, $x = 0.6$, we evaluate (C.6) with $h = 1$, thresholds $K \in \{1, 2, 5, 10, 20\}$ and loads $\lambda \in \{0.5, 2.8, 3.4\}$. These 15 scenarios specify the comparison grid; no global search over policies or integer thresholds is claimed. $K = 1$ is immediate renewal. Each finite threshold has the same capacity guarantee from Theorem C.5, but their delays differ.

Threshold K	$\lambda = 0.5$	$\lambda = 2.8$	$\lambda = 3.4$
1	0.6703	2.8120	8.0000
2	0.3811	2.7844	8.0000
5	0.2305	2.6541	8.0670
10	0.2289	2.5923	8.5127
20	0.2289	2.9386	10.3677

Table 3. Mean authorisation delay in weeks under backlog-triggered renewal. Values come from the unbounded stationary calculation; all scenarios are constructed.

At $\lambda = 2.8$, $K = 10$ lowers mean delay from 2.8120 under immediate renewal to 2.5923, about 7.8 per cent. The same threshold gives 8.5127 weeks at $\lambda = 3.4$, compared with 8.0000 under immediate renewal. At $\lambda = 0.5$, it nearly matches the case-first mean in the displayed precision. This is evidence about the specified numerical model, not a theorem of low-load dominance or convergence as K grows. A finite threshold avoids the above-capacity starvation of strict case-first review but does not provide a workload-independent delay improvement.

The first implementation solves the finite boundary levels $0, \ldots, K$ joined to an unbounded homogeneous matrix-geometric tail. The second constructs a finite event generator directly, with queue cutoffs of 400 and 800. At cutoff 800 every mean agrees with the unbounded calculation to within 10^{-7} weeks and boundary mass is below 10^{-10}. Both methods check their balance residuals and probability normalisation. These tests address truncation and implementation, not uncertainty in the model assumptions.

The same separate event simulator evaluates all 15 scenarios with 20 independently seeded runs each, using the same 10,000-unit warm-up and 100,000-unit measurement duration as the earlier checks. Thirteen of the 15 pointwise approximate 95 per cent intervals contain the stationary mean. The exceptions are $\lambda = 2.8$, $K = 20$ (simulation 2.9079, interval $[2.8834, 2.9324]$, stationary mean 2.9386) and $\lambda = 3.4$, $K = 5$ (simulation 7.7781, interval $[7.5011, 8.0551]$, stationary mean 8.0670). Both misses are reported without replacing seeds or selectively extending runs; pointwise intervals are not simultaneous guarantees. The full table, run-level results and source hashes are supplied. The matrix agreement is the more precise numerical check.

For the uncertainty certificate, 1,000 independently sampled parameter rectangles and allocations are checked at their 32 vertices, giving 32,000 capacity checks. The 99 sampled boxes satisfying the sufficient inequality additionally give 3,168 common-drift checks. The proof in Appendix C.9 establishes the result for the whole stated set; these sampled checks only test the formula's implementation. No field parameter intervals are inferred from these synthetic checks.

E Empirical design and institutional scope

E.1 A conditional allocation decision

First establish that a standing policy and a manual fallback are admissible. Then estimate the rate of incoming candidates, eligible share, valid-policy lifetimes, active renewal effort and case-review effort. Record policy invalidation separately from the time at which renewal starts. Calendar renewal delay includes queueing; it cannot be substituted for full-effort renewal time $1/\nu$.

If λ is safely below μ, ordinary assisted review is a viable capacity baseline. Theorem C.4 and equation (D.2) show why neither immediate renewal nor permanent manual review should automatically be selected for delay. If λ exceeds μ but is below the renewal frontier, a scheme that renews only when no cases remain has no stability guarantee and, in this model, fails. A protected renewal share can restore stability. Theorem C.5 permits case priority below a fixed finite backlog threshold; Proposition C.1 supplies a worst-corner certificate when fixed parameters are uncertain. An operational choice still needs additional slack and a delay target. If λ exceeds λ_{cap}, merely rearranging the same reviewer cannot stabilise the model. Demand, eligibility, review or renewal productivity, validity duration, or available capacity must change.

These are conditional implications of measured parameters, not instructions to delegate whenever $\nu x > \delta$. The frontier concerns retained cases and admissible authorisation; harmful approvals can be fast. A slower valid rejection may have greater value than a faster harmful release. Setup, maintenance tooling, monitoring, interruptions and specialist staffing also have costs outside the queue objective.

E.2 A test that distinguishes the mechanism

A prospective study should compare assisted case review, case-first renewal and protected renewal for a prespecified class with a common assistance configuration and total reviewer budget. Use a documented workflow in which policy validity is independently auditable and both individual review and policy renewal are permitted. DORA's approval measures provide a practical starting point for software changes [7]; they do not validate the model's exact assumptions.

Randomise at the shared review-resource level where feasible. Cases using the same reviewer are not independent treatment units because one class's renewal can change the other's wait. If a phased design is necessary, identify time trends, learning and concurrent reforms. Analyse channels by assigned condition and report actual allocation as a mechanism measure. A before-and-after change alone does not isolate the effect of protecting renewal time.

Estimate μ, ν, x and δ from an initial period and predict the load range and reservation threshold for a later period. Record arrivals, policy validity, allocated effort, rework, queue occupancy and all dispositions. The key prediction is not merely that delegation helps. It is that case-first renewal becomes unreliable above measured ordinary capacity, while a sufficient protected share maintains policy availability and prevents sustained backlog growth. Below ordinary capacity, the direction of the delay comparison can differ.

A safety or validity restriction must never be deliberately breached to test instability. Use existing variation, retrospective logs or a shadow replay where overload cannot cause harm. A finite field interval cannot prove positive recurrence. Assess bounded service-level outcomes and predictive performance over the observation horizon instead of equating a short period without backlog growth with asymptotic stability.

E.3 Quality and institutional scope

Define a candidate cohort upstream of treatment-dependent triage, include rejected, cancelled and unfinished work, and assess eligibility under common criteria. Measure useful completion, valid rejection, harmful authorisation, missed escalation and policy-maintenance effort. Use independent outcome assessment and a prespecified acceptable quality-loss margin. Failure to detect a significant quality difference is not evidence of equivalence. Where treatment changes case generation, report outcomes per fixed resource-time as well as cohort proportions.

The EU AI Act's Article 14 concerns effective oversight of high-risk AI systems and does not by itself prescribe one universal case-review schedule [8]. Banking SR 26-2 is a separate, risk-based governance reference; its scope excludes

generative and agentic AI [4, 5]. These sources motivate careful definition of admissible controls. They do not certify the model's routing choices.

For DSI, the testable implication is that a system should make policy status and renewal workload observable and support protected allocation where needed. That is an implementation hypothesis, not evidence that a proprietary platform is necessary. Ordinary workflow tools can implement the same rule. The organisational value must be demonstrated through outcomes and full costs.

E.4 Estimating the quantities the certificate requires

The model can be fitted to a complete event log under its stated observation assumptions. Over a fixed observation window of duration T, let N_A count all candidate arrivals, including immediate authorisations; N_S count manual completions; N_R count completed renewals; and N_I count invalidations. Let E_S be actual effort spent serving manual cases, E_R effort spent renewing while invalid, and T_V observed valid-policy time. Let N_V count arrivals during valid periods and N_B the eligible cases among them. The model likelihood, conditional on the initial state and a known predictable allocation rule, contains the terms

$$\begin{aligned}\ell = {} & N_A \log \lambda - \lambda T + N_S \log \mu - \mu E_S \\ & + N_R \log \nu - \nu E_R + N_I \log \delta - \delta T_V \\ & + N_B \log x + (N_V - N_B) \log(1 - x) + \text{constant}.\end{aligned} \tag{E.1}$$

The corresponding interior maximum-likelihood estimates are

$$\widehat{\lambda} = N_A/T, \quad \widehat{\mu} = N_S/E_S, \quad \widehat{\nu} = N_R/E_R, \quad \widehat{\delta} = N_I/T_V, \quad \widehat{x} = N_B/N_V. \tag{E.2}$$

These follow from the event intensities in Appendix B and independent eligibility marks. Exposure includes unfinished service and renewal at the end of the observation window; completion-only duration averages would omit that censoring. E_S excludes idle time and E_R excludes waiting to start renewal. Without valid-policy arrivals, x is not identified by this log; without the relevant exposure or event counts, a positive interior rate estimate may not exist. Such cases must not be converted into a confident capacity certificate.

Fit on an initial window, assess arrival and duration assumptions, and evaluate predictions on held-out windows. Construct a joint uncertainty region with a coverage method justified for the observation design and dependence, then use a conservative rectangular enclosure if desired. Separate marginal 95 per cent intervals do not automatically produce a joint 95 per cent certificate. The supplementary logging specification provides the required fields and accounting checks. This is an estimation protocol derived from the model; no dataset has been fitted in this paper.

F General renewal durations

F.1 Statement and scope

The stationary delay formula and the case-first hitting probability rely on the Markov model. The immediate-renewal capacity condition has a broader renewal interpretation.

Proposition F.1. *Under immediate renewal, replace exponential valid lifetimes and renewal effort by positive cycle pairs (U, R) that are independent and identically distributed across cycles, independent of arrivals and case requirements, with finite means $E[U] = u$ and $E[R] = r$. Here U is calendar time valid and R is full-effort renewal time. Cases have independent and identically distributed non-negative work requirements with finite mean $1/\mu$, independent of arrivals, eligibility marks and cycle pairs. With the same independent eligible share x, the stationary reflected case-workload process exists under:*

$$\frac{\lambda[(1 - x)u + r]}{\mu} < u. \tag{F.1}$$

Strict reversal of the inequality gives positive long-run net input and rules out a proper stationary workload. Equality is not classified here. With $u = 1/\delta$ and $r = 1/\nu$, equation (F.1) reproduces the immediate-renewal boundary in Theorem C.1. This proposition concerns that policy's workload stability; it does not extend every Markov control or delay result to arbitrary duration laws.

During a valid period the reviewer can process case work for U time units, while renewal consumes the whole reviewer for R. Expected new manual work per cycle is $\lambda[(1 - x)u + r]/\mu$. The reflected-input construction and regenerative strong law establish the result; the argument is set out in Appendix F.2 using established stationary-workload reasoning [16]. Finite first moments suffice for a proper stationary workload under strict negative drift. They do not guarantee finite mean delay. In particular, long renewal tails can produce severe delays without changing the average capacity boundary.

The model also assumes that manual service remains available while the policy is invalid. If policy invalidation also removes the legal or operational basis for individual review, strict case-first service is not an admissible fallback and the

control problem changes. If completed renewal can release the existing backlog, or if renewal is conducted by a separate resource, both the transition structure and capacity bound must be changed.

F.2 Proof for general renewal durations

Consider the policy-alternating environment in its stationary version, which exists because $u + r$ is finite. Under immediate renewal its transitions do not depend on the queue. Let $B(s,t)$ be the manual case-work arriving during $(s,t]$, and let $T_{\text{valid}}(s,t)$ be the time available for case service there. The queue is the reflection at zero of net input $X(s,t) = B(s,t) - T_{\text{valid}}(s,t)$. Regenerative strong laws give long-run net input per unit time:

$$g = \frac{\lambda[(1-x)u + r]/\mu - u}{u + r}. \tag{F.2}$$

If $g < 0$, $X(-t,0)$ tends to minus infinity as t grows. Local finiteness of arrivals and finite case requirements make $\sup\{X(s,0) : s \le 0\}$ finite almost surely. That supremum constructs a proper stationary workload by reflection, following the standard stationary-input argument [16]. If $g > 0$, forward net input grows linearly and reflection cannot remove positive drift, ruling out a proper stationary workload. No finite expected workload follows from this construction without further moment conditions. The result permits dependence between U and R within each independent cycle pair, but not dependence of those pairs on the realised queue or service requirements.

G Reproducibility

The accompanying source package contains the LaTeX manuscript, the header artwork and vector mechanism figure, two stationary calculations, a separate C++ event simulator, run-level outputs, a duration-sensitivity calculation and an event-log specification. The Python scripts require NumPy and SciPy; the simulator requires a C++17 compiler. The README records commands, initial conditions, seeds, versions and the interpretation of each output. A checksum manifest identifies the source and result files.

All reported numerical results are generated by the supplied implementations of the stated transition model. The package records the full scenario grid, run-level simulation estimates and intervals, finite-cutoff comparisons and unbounded stationary solutions. It also states the limitations of pointwise Monte Carlo intervals and retains every reported run without selecting seeds to improve agreement.

The finite generator and matrix-geometric calculations are separate numerical formulations, and the event simulator uses neither stationary solution. This checks agreement between implementations; it is not independent empirical replication. The supplied computations contain constructed cases only. No operational logs, human outcomes or organisational treatment effects are generated or represented by them.

References

[1] Philippe Aghion and Jean Tirole. Formal and Real Authority in Organizations. *Journal of Political Economy*, 105(1):1–29, 1997. doi:10.1086/262063.

[2] Søren Asmussen. *Applied Probability and Queues*. Springer, New York, second edition, 2003. doi:10.1007/b97236.

[3] Susan Athey, Joshua Gans, Scott Schaefer, and Scott Stern. The Allocation of Decisions in Organizations. Working Paper 1322, Stanford Graduate School of Business, 1994. URL https://www.gsb.stanford.edu/faculty-research/working-papers/allocation-decisions-organizations.

[4] Board of Governors of the Federal Reserve System. SR 26-2: Revised Guidance on Model Risk Management, April 2026. URL https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm. Supervisory letter, 17 April 2026.

[5] Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, and Federal Deposit Insurance Corporation. Supervisory Guidance on Model Risk Management, April 2026. URL https://www.federalreserve.gov/frrs/guidance/supervisory-guidance-on-model-risk-management.htm. Interagency guidance, 17 April 2026; Purpose and Scope, note 3.

[6] Decision Superintelligence Labs. Enterprise AI Adoption as an Organisational State Transition. Research paper, August 2026. URL https://www.dsilabs.ai/research/enterprise-ai-adoption-state-transition. Dated 1 August 2026.

[7] DORA. Streamlining change approval, October 2025. URL https://dora.dev/capabilities/streamlining-change-approval/. Research and capability guidance, updated 30 October 2025.

[8] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024. *Official Journal of the European Union*, L, 2024/1689, July 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Article 14: Human oversight.

[9] Awi Federgruen and Kut C. So. Optimal time to repair a broken server. *Advances in Applied Probability*, 21(2):376–397, 1989. doi:10.2307/1427165.

[10] R. Kalyanaraman and A. Sundaramoorthy. A Markovian Working Vacation Queue with Server State Dependent Arrival Rate and with Unreliable Server. *International Journal of Computer Applications*, 179(39):1–7, 2018. doi:10.5120/ijca2018916926.

[11] David L. Kaufman and Mark E. Lewis. Machine maintenance with workload considerations. *Naval Research Logistics*, 54(7):750–766, 2007. doi:10.1002/nav.20248. Author manuscript: https://public.websites.umich.edu/~davidlk/pubs/reliability.pdf.

[12] Junji Koyanagi and Hajime Kawai. An optimal maintenance policy for a server with decreasing arrival rate. *Mathematical and Computer Modelling*, 31(10–12):225–234, 2000. doi:10.1016/S0895-7177(00)00091-1.

[13] Ravi Kumar, Mark E. Lewis, and Huseyin Topaloglu. Dynamic service rate control for a single-server queue with Markov-modulated arrivals. *Naval Research Logistics*, 60(8):661–677, 2013. doi:10.1002/nav.21560.

[14] Michael Lin, Richard J. La, and Nuno C. Martins. Stabilizing a Queue Subject to Activity-Dependent Server Performance. *IEEE Transactions on Control of Network Systems*, 8(4):1579–1591, 2021. doi:10.1109/TCNS.2021.3074527. Author version consulted: arXiv:1903.00135v3.

[15] Ruihan Lin and Jiheng Zhang. When to Screen, When to Bypass: LLM-Judges in Resource-Scarce AI-Human Workflow. arXiv:2603.13870v1, March 2026. Version consulted submitted 14 March 2026.

[16] R. M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Proceedings of the Cambridge Philosophical Society*, 58(3):497–520, 1962. doi:10.1017/S0305004100036781.

[17] Thodoris Lykouris and Wentao Weng. Learning to Defer in Congested Systems: The AI-Human Interplay. arXiv:2402.12237v4, August 2025. First submitted 19 February 2024; version consulted revised 12 August 2025.

[18] Andrew W. Shogan. A single server queue with arrival rate dependent on server breakdowns. *Naval Research Logistics Quarterly*, 26(3):487–497, 1979. doi:10.1002/nav.3800260310.