

**When AI Output Outpaces Human Review:
A Capacity-Aware Framework for Enterprise Automation**

Sruthi Rachamalla

Southern Illinois University, Carbondale

ORCID iD: 0000-0001-9312-507X

Abstract

Background: Faster AI output creates little value when the people responsible for checking it cannot keep pace. Review capacity must therefore influence how much work an organization initiates, not merely how it processes completed outputs.

Methods: This conceptual paper draws on research in workplace AI, human decision-making, queueing, and risk management to develop an operational framework for review-dependent workflows. It combines reviewer-time reservations, consumed-time accounting, skill-based routing, explicit deferral, and quality feedback.

Results: The proposed architecture admits work against a bounded review budget and preserves mandatory checks during congestion. A fictional customer-service example shows how capacity limits expose an unresolved demand gap instead of transferring it invisibly to reviewers. A decision table specifies responses to overload, missing expertise, and deteriorating review quality.

Conclusion: The appropriate production rate depends on the organization's ability to verify outcomes. The framework offers an implementation starting point, not a validated performance claim. Its effectiveness requires testing under variable arrivals, correction workloads, and reviewer availability.

Keywords: Human oversight; enterprise AI; review capacity; admission control; workload management

1 Introduction

Suppose an AI assistant prepares fifty customer responses while a supervisor checks five. The extra output looks productive, but customers receive no benefit until the responses are verified and released. The practical question is how much unfinished work the organization should allow.

Evidence supports both opportunity and caution. Brynjolfsson, Li, and Raymond found a 15% average productivity increase among customer-support agents, with substantial differences across workers [1]. A survey of 319 knowledge workers associated greater confidence in generative AI with less reported critical thinking; it did not establish causation [2]. Bartolucci and Vivo's queueing preprint explicitly examines congestion and rework in AI-assisted workflows [3].

The framework turns review constraints into operational admission rules through task-level reservations, reviewer eligibility checks, and visible deferral.

2 Methods

The framework is a targeted design synthesis, not a systematic review or deployment study. NIST's AI Risk Management Framework supplies the foundation for defined oversight responsibilities [4]. Published human-factors evidence informs review design; queueing research motivates workload accounting [2], [3].

The scope is administrative work requiring approval before release. For each task, record its owner, arrival time, deadline, required reviewer skills, evidence, and estimated review effort. Include correction, repeat review, and escalation in that estimate. Estimates come from observed comparable tasks; unfamiliar work receives a conservative allowance.

Capacity excludes meetings, breaks, and other duties. Reviewers must not be counted in multiple pools simultaneously. The unit of capacity is qualified reviewer time, not employee headcount. Available generalists cannot substitute for a specialist required to release a particular case.

3 Results

3.1 Reserve capacity before generation

Figure 1 places an admission gate before AI preparation. For each reviewer pool and planning window, C_{window} is total usable reviewer-minutes, S is time already consumed, B is remaining reserved effort, and e is estimated effort for a new task. The gate applies:

$$S + B + e \leq \alpha C_{\text{window}} \quad (1)$$

Here, α is chosen locally between zero and one to preserve spare capacity. The rule is a budgeting heuristic, not a guarantee of queue stability or deadlines. Admission also requires a feasible assignment before the task's deadline. Reservations are atomic, preventing parallel agents from spending the same capacity. A task requiring two specialties reserves both pools before preparation. If either reviewer becomes unavailable, the task is reassigned or held. Spare capacity in a routine queue cannot substitute for missing specialist capacity.

Each task retains one reservation through preparation, waiting, review, and corrections. As review proceeds, consumed time moves from B to S . Closing or canceling a reservation releases only unused time; it never restores consumed capacity. Update effort estimates after preparation; overruns reduce new admissions, not required checks. Carry unfinished effort into the next window. Decisions use approved staffing and policy records, not model confidence. Every deferral retains its reason and original arrival time.

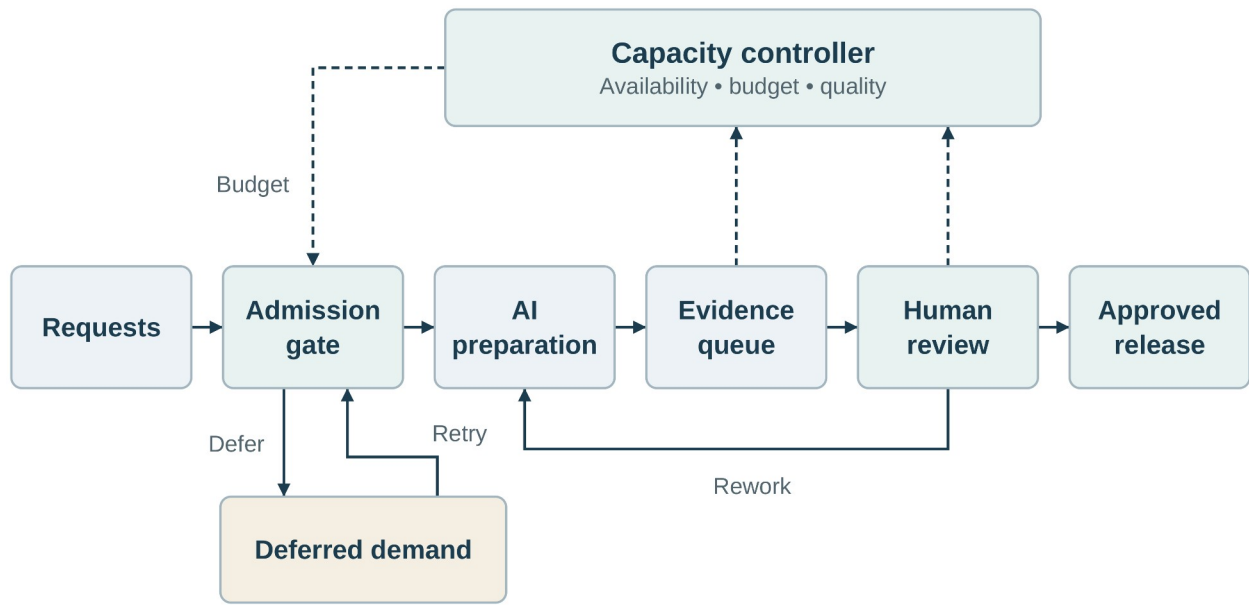


Figure 1. Review capacity governs admission; feedback adjusts the budget without bypassing human approval.

3.2 Protect the review itself

Each output arrives with its sources, applicable policy, unresolved assumptions, and proposed business effect. Reviewers can inspect the evidence, reject the proposal, or request clarification. Experiments on cognitive forcing suggest that deliberate scrutiny can reduce overreliance, although effective interventions may be less convenient [5].

Tasks are assigned only to qualified reviewers. Prioritization considers consequences and deadlines, with age limits preventing routine requests from waiting indefinitely. Table 1 summarizes control responses. Queue pressure never authorizes automatic approval of work requiring review. Sampling may supplement oversight of eligible routine work, but cannot replace a mandated review. Reviewers can decline unfamiliar cases without being penalized for lowering throughput.

Table 1. Operating responses; thresholds require local validation.

Condition	Operational response
Capacity available	Reserve time; admit eligible work.
Budget exhausted	Defer generation; retain arrival time and owner.
Expert unavailable or deadline infeasible	Escalate; add qualified capacity or revise commitment.
Review errors increase	Reduce admissions; investigate; preserve required checks.

3.3 Illustrative capacity calculation

Assume four reviewers each provide 45 review-minutes per hour. Capacity is 180 minutes. Each case requires six minutes, including expected corrections. Unrestricted generation of 50 cases therefore creates 300 minutes of hourly demand: a 120-minute deficit under these assumptions.

With $\alpha = 0.80$, no consumed time, and no carryover, the initial budget admits 24 cases: 144 minutes of expected work, leaving 36 minutes of spare capacity. The remaining 26 requests stay visible upstream with their original timestamps. They are not completed cases or eliminated demand. Sustained arrivals require added qualified capacity, reduced demand, or an explicitly changed service commitment. The calculation illustrates budgeting; it is not a measured improvement or completion forecast.

4 Discussion

Task reservations translate workload estimates into admission decisions. Delayed generation can protect review conditions, but cannot create human capacity. This operational design complements analytical work on congestion [3].

Evaluation should compare unrestricted generation, a fixed work-in-progress cap, and the proposed gate using the same requests and reviewer hours. Test burst arrivals, absences, underestimated effort, and difficult cases. Independent adjudication should assess consequential errors missed during review. Report verified completions, time from initial request to release, deadline failures, correction effort, and all deferred or abandoned requests. Reviewer workload must accompany throughput. Report both upstream waiting and downstream review time: a shorter review queue is not an improvement if the same delay has merely moved elsewhere.

A small pilot should establish baseline review times before enforcement. Separate pause and restart thresholds can prevent repeated switching. Check estimates by case type, not only overall averages. Rare, lengthy corrections can consume spare capacity; repeated underestimation should reduce admissions until staffing or estimates are corrected. Persistent errors should tighten admission and trigger investigation, not faster approvals. Inaccurate estimates, biased prioritization, and skill shortages remain limitations; capacity monitoring does not replace sound authorization or competent judgment. Early use should remain advisory until estimates are credible. Use workload records for planning, not to rank individuals by approval speed, which could reward superficial checking.

5 Conclusion

Enterprise AI should produce work at a pace the organization can responsibly verify. Reserving reviewer time, preserving evidence, and exposing deferred demand make oversight operational. Success is a defensible outcome delivered within a realistic service commitment—not a growing pile of unchecked output.

References

- [1] E. Brynjolfsson, D. Li, and L. Raymond, “Generative AI at work,” *The Quarterly Journal of Economics*, vol. 140, no. 2, pp. 889–942, 2025, doi: 10.1093/qje/qjae044.
- [2] H.-P. Lee et al., “The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers,” in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2025, Art. no. 1121, pp. 1–22, doi: 10.1145/3706598.3713778.
- [3] S. Bartolucci and P. Vivo, “Queue & AI: When faster tasks slow down the workflow,” *arXiv:2605.27202*, 2026, preprint, doi: 10.48550/arXiv.2605.27202.
- [4] E. Tabassi, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, 2023, doi: 10.6028/NIST.AI.100-1.
- [5] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, “To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making,” *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, Art. no. 188, pp. 1–21, 2021, doi: 10.1145/3449287.