

# Beyond the Model Frontier

Acceptance Boundaries and Organizational Scaling in the AI Race

J. D. Miller

August 21, 2026

## Abstract

The race toward increasingly general AI is usually measured upstream: compute, benchmarks, model prices, capital, and task horizons. Organizations experience it downstream, where generated candidates meet identity, permissions, data rights, review, production state, commitments, and recovery. This paper develops the *acceptance boundary*—the governed passage from an AI-generated candidate to an authoritative organizational effect—and conducts a structured integrative review of causal studies, preprints, filings, standards, telemetry, and named enterprise and venture-capital cases through August 21, 2026.

The evidence does not support a universal productivity effect. Bounded tasks show substantial gains, yet coding-to-release, generation-to-review, and signal-to-screening evidence shows attenuation. This pattern is consistent with candidate production outrunning shared capacity to validate, authorize, integrate, observe, and recover state while transferring exception load to people without protected recovery time. We distinguish observed acceptance, contemporaneous validity, and delayed value; model valid-acceptance throughput and fully loaded cost; represent an autonomy sweet spot as a consequence-constrained Pareto region; and introduce a proposed acceptance-capacity runway intended to estimate, under explicit scenarios, whether an operating boundary may bind before new capacity is ready.

The paper contributes a seven-boundary stack, a model of boundary load and acceptance-capacity runway, a behavior-bundle release unit, and a four-plane reference architecture designed to make shared control functions reusable while preserving federated workflow ownership, together with nine falsifiable propositions. Cases are consistent with scalable configurations combining high candidate volume with bounded authority, measurable postconditions, narrow tools, independent evaluation, continuity, and recovery. Venture evidence adds that AI can widen sourcing and synthesis while reducing the diagnostic value of polished narratives. Agnostic about AGI timelines, we argue that once configured capability clears a task threshold, the institutional race shifts toward scaling valid acceptance, revocable authority, and recoverable state before shared boundary capacity binds.

**Keywords:** enterprise AI; agentic systems; organizational scaling; acceptance boundary; AI infrastructure; observability; cybersecurity; psychosocial risk; venture capital; human-AI collaboration; AGI

**Research status.** This is a structured integrative review, not a systematic review, meta-analysis, or causal evaluation of the proposed framework. Peer-reviewed findings, working papers, company-reported cases, and conference demonstrations are labeled separately. The runway measure and conditional operating states are author-generated decision aids. Scenario essays about rapid AGI development are used to motivate a stress test, not to establish a timeline.

# 1 Introduction: the race meets the institution

Scaling-law research documented empirical power-law regularities for language-model loss over studied regimes, while compute-optimal work showed that model size, data, and compute must be balanced rather than enlarged independently [Kaplan et al., 2020, Hoffmann et al., 2022]. Inference-time reasoning and search add another scaling axis on studied reasoning benchmarks [Snell et al., 2024]. None of these results guarantees general capability or operational reliability. They help explain a race organized around training clusters, data, algorithms, model releases, and increasingly agentic interfaces. Scenario reports extend that trajectory into compressed paths toward broadly capable or superhuman systems [Aschenbrenner, 2024, Kokotajlo et al., 2025]. Their timelines are contested and are not adjudicated here.

The organizational problem begins where the frontier framing ends. A firm does not deploy a benchmark score. It deploys a socio-technical system inside real entitlements, service dependencies, budgets, data obligations, incentives, and records. A model may draft a credit analysis, a customer response, an investment memo, or a code change. The institution creates an authoritative record, commitment, or state change only when some result is accepted into a system of record, sent externally, used to allocate capital, or allowed to alter rights and resources. Acceptance does not make the content true. Fluency is therefore neither necessary nor sufficient evidence of organizational success. Authority, trajectory, evidence, postcondition, and recovery matter as well.

This paper defines the governed passage from generated candidate to authoritative effect as an **acceptance boundary**. The phrase itself is not claimed as a coinage. The construct is deliberately narrower than organizational accountability boundaries, which locate responsibility and signoff across firms and complementors [Hydari and Muzaffar, 2026]. Here the unit is a runtime event inside a workflow: pre-action authority, execution through permitted systems, and post-action validation. A draft that never leaves a sandbox may stop before the boundary. A message to a customer, update to a ledger, production deployment, approval of a claim, or funds movement crosses it. Many workflows cross several such boundaries.

The distinction has become empirically visible. In randomized field experiments, coding assistants increased completed tasks by about 26 percent across 4,867 developers at Microsoft, Accenture, and an anonymous Fortune 100 company [Cui et al., 2026]. In a working paper linking telemetry from more than 100,000 GitHub developers to successive generations of coding tools, estimated cumulative effects rose as high as 180 percent for commits but attenuated to 50 percent for projects and 30 percent for releases; app-marketplace evidence showed more new apps but no increase in total usage [Demirer et al., 2026]. In a separate longitudinal case, an AI-forward software company more than doubled merged pull requests per developer while per-reviewer load also roughly doubled and substantive human review thinned [He et al., 2026]. These studies differ in design and cannot be pooled as one coefficient. Together they motivate a production hierarchy in which upstream acceleration can meet downstream weak links; task composition, selection into use, demand, and release strategy remain competing explanations for the observed attenuation.

The same issue appears outside software. A randomized experiment with 758 consultants found that GPT-4 improved speed, task completion, and quality for tasks inside its tested frontier, but reduced correctness on a selected task outside that frontier [Dell’Acqua et al., 2026b]. A field experiment with 791 Procter & Gamble professionals found that individuals with AI matched teams without AI on product-innovation work, while human judgment retained value in selecting among ideas [Dell’Acqua et al., 2026a]. In venture and hiring experiments, giving senders access to ChatGPT improved writing but made the resulting materials less diagnostic of underlying expertise on average, lowering screening accuracy by 4–9 percent and increasing demand for costlier corroborating signals by 6–9 percent

[Cowgill et al., 2026]. The common object is not generated output. It is accepted, decision-relevant work.

The paper asks five research questions:

- RQ1:** What boundaries separate model capability from organizationally valid acceptance and delayed value?
- RQ2:** Under what task and system conditions does increasing autonomy improve the valid-value, cost, loss, and continuity frontier rather than merely candidate volume?
- RQ3:** Which infrastructure, identity, data, security, observability, and recovery functions should be shared across workflows, and which outcome and residual-risk decisions must remain with accountable domain teams?
- RQ4:** Which leading indicators can forecast when those shared operating functions will bind before new capacity can be safely made ready?
- RQ5:** What do enterprise, startup, and venture-capital cases establish – and fail to establish – about those conditions?

The contribution is a synthesis across literatures that often stop at different layers. Model research evaluates capability; human-computer interaction studies assisted performance; economics estimates productivity and adoption; security research tests exploit mechanisms; reliability practice manages services; and management research examines organizational complements. The acceptance boundary supplies one common outcome unit. The paper then adds five analytical artifacts:

- a **boundary stack** that distinguishes capability, epistemic, authority, systems, organizational, recovery, and capital constraints;
- a **boundary-load and runway model** connecting candidate arrival rate, validation effort, acceptance capacity, build lead time, expected loss, and accepted throughput, together with a consequence-constrained autonomy region;
- a versioned **behavior-bundle manifest** that records the deployed model, data, tools, permissions, policies, evaluators, and recovery configuration;
- a **four-plane operating architecture** that separates reusable shared control, evidence, and technical-recovery services from federated, domain-owned workflow execution and outcome responsibility; and
- nine propositions with an operational test and falsifier.

The central thesis is conditional: model capability remains decisive until it clears the threshold for a given task. After that threshold is crossed, the institution’s capacity to validate, authorize, integrate, observe, and recover can become the binding constraint. The operational implication is not a permanent manual-review bureaucracy. It is a shared operating substrate—with named infrastructure, data, platform, security, and reliability owners—that makes evidence, authority, and technical recovery machine-scalable while domain teams retain workflow meaning, postconditions, exception ownership, and residual-risk decisions. A sufficiently capable and reliably self-verifying system could reduce the boundary burden. So could better workflow design, stronger platforms, or removal of obsolete approvals. The framework is useful only if its load and runway measures predict where value survives contact with real operations better than model, seat, or adoption metrics alone. Its forward claim is operational rather than an AGI timetable: which boundary resource will lose safe headroom before replacement capacity can be made ready.

The paper proceeds from construct to consequence. Sections 2–3 define the evidence rules, boundary stack, accepted-outcome measures, load, runway, autonomy region, and behavior bundle. Sections 4–5 examine gains, attenuation, human recovery, organizational complements, and the emerging sweet spot. Section 6 specifies the shared operating substrate, its security and remediation obligations, and the conditional capacity forecast. Section 7 applies the framework to venture capital and startups;

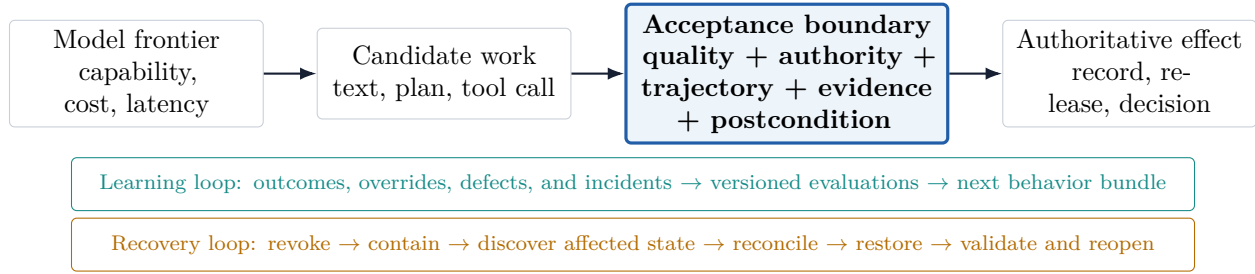


**Figure 1:** The analytical shift from model output to accepted organizational outcome. Learning and recovery are separate operating loops; neither implies that all failures are reversible.

Sections 8–10 state falsifiable propositions, operating implications, limitations, and a research agenda.

## 2 Method and evidence discipline

### 2.1 Structured integrative review

The review covers material available through August 21, 2026. Searches were organized around five domains: (1) model scaling, evaluation, and agent reliability; (2) workplace productivity and organizational adoption; (3) infrastructure, identity, security, observability, and recovery; (4) venture operations, startup organization, and AI investment; and (5) named deployment cases with enough detail to locate an acceptance boundary. Sources were discovered through scholarly indexes, arXiv, official research repositories, regulator and standards sites, public-company filings, conference archives, and the source trails of retained papers. User-provided reports and documents informed topic selection but were not treated as evidence merely because they were provided.

The review is purposive rather than exhaustive. It is designed to test and refine an analytical framework, not to estimate a pooled treatment effect. Sources were retained when they supplied at least one of the following: a causal or longitudinal design; a transparent sample and outcome definition; an authoritative statement of deployed architecture or risk; a documented exploit or failure mechanism; a public filing; or a named case that located both the automated task and the retained decision right. Pure forecasts, unsupported productivity claims, and vendor case studies without a discernible workflow were excluded from the core evidence. Scenario work was retained only as frontier context. The runway measure and conditional operating states are author-generated inferences from the retained evidence, not retained forecast evidence or predictions of model capability or AGI timing.

### 2.2 Source types and claim rules

Table 1 summarizes the source types. They are descriptive rather than ordinal: a filing is stronger than an experiment for establishing what a company disclosed, while an experiment is stronger for causal performance. The rule throughout is to match the claim to the source.

Every major empirical statement is interpreted using four questions: What does the source directly measure? Where does the inference stop? Which boundary mechanism is consistent with the result? What observation would weaken the proposed explanation? The analysis keeps the following distinctions explicit:

- capability versus operating authorization;
- generated candidates, observed acceptance, contemporaneous validity, and delayed value;
- pass-at-one versus repeated reliability;

**Table 1:** Source types used in the integrative review

| Source type          | Examples                                                                                                         | What it can establish                                                                    | Principal limitation                                               |
|----------------------|------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|--------------------------------------------------------------------|
| CAUSAL STUDY         | Peer-reviewed randomized, quasi-experimental, or controlled field research                                       | Causal or strongly identified effects in the studied setting                             | Task, model, worker, time, and site specificity; external validity |
| AUTHORITATIVE RECORD | Public filings, regulators, government data, standards, maintainer advisories, operational telemetry with method | Disclosed practice, market facts, formal guidance, vulnerability or usage pattern        | Often descriptive; incentives and definitions vary                 |
| SCHOLARLY EMPIRICAL  | Peer-reviewed benchmarks and observational studies; transparent working papers and preprints                     | Reproducible behavior, mechanisms, associations, longitudinal patterns, or formal models | Transfer and causal limits; preprint results may change            |
| DEPLOYMENT REPORT    | Vendor or company case studies, conference demonstrations, practitioner surveys                                  | Feasibility, implementation pattern, directional evidence                                | Selection, marketing, self-report, and missing counterfactuals     |

- final-answer correctness versus a policy-conforming trajectory;
- product access or usage versus organizational value;
- exploit feasibility versus production prevalence;
- financing versus valuation, and valuation versus liquidity;
- technical failover versus state-aware recovery; and
- technical recovery of authority and state versus protected recovery opportunity for workers.

Named company cases are treated as comparative observations, not representative prevalence estimates. Morgan Stanley, BBVA, SignalFire, and OpenAI case pages are useful for locating task boundaries, but vendor-authored numbers remain self-reported [OpenAI, 2025, 2026a, Kirnos, 2023]. Klarna’s public filing is more authoritative about what management disclosed, but its efficiency and quality measures still depend on internal methods [Klarna Group plc, 2026]. The anonymous enterprise coding study offers rich telemetry but one unusually favorable site and a non-random rollout [He et al., 2026]. No case is allowed to carry the general thesis alone.

### 2.3 Positioning and framework crosswalk

The individual mechanisms assembled here are not new discoveries. Prior work already locates accountability across organizational boundaries, separates stages and levels of automation, documents a jagged capability frontier, measures information-technology complements, studies automation bias and out-of-the-loop performance, and evaluates models and agents. The proposed novelty is a cross-level synthesis and operationalization: it makes the governed state transition the common unit, joins shared acceptance capacity to authority and recovery, and carries the construct into enterprise and venture settings. Table 2 states that boundary explicitly.

Table 3 prevents the framework’s components from being read as competing taxonomies. The boundary stack describes antecedent constraints; the load model describes shared acceptance capacity; the manifest defines the release and evidence unit; the four planes specify proposed implementation functions; and the human and capital applications introduce continuity, delayed truth, selection,

**Table 2:** Related work and the extension claimed here

| Literature                                                                                                                                                  | Primary unit or outcome                                           | Enterprise variable left open                                 | Extension in this paper                                                     |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------|---------------------------------------------------------------|-----------------------------------------------------------------------------|
| Accountability boundaries [Hydari and Muzaffar, 2026]                                                                                                       | Responsibility and signoff across firms and complementors         | Runtime passage from candidate to authoritative state         | Event-level authority, trajectory, evidence, postcondition, and recovery    |
| Levels and stages of automation [Parasuraman et al., 2000]                                                                                                  | Allocation of acquisition, analysis, selection, and action        | Shared validation load and local production evidence          | Multidimensional authority profiles tied to valid acceptance                |
| Jagged frontier and human-AI work [Dell’Acqua et al., 2026b,a]                                                                                              | Performance on bounded tasks and team configurations              | Production queues, permissions, state integrity, and recovery | Task fit joined to consequence-specific operating constraints               |
| Model and agent evaluation [Liang et al., 2023, Yao et al., 2025, Kapoor et al., 2025, Cemri et al., 2025]                                                  | Accuracy, robustness, cost, trajectories, or system failure modes | Organizational authorization and delayed business value       | Versioned behavior bundles and accepted-outcome denominators                |
| IT complements and productivity J-curve [Bresnahan et al., 2002, Brynjolfsson et al., 2021]                                                                 | Firm productivity and complementary intangible investment         | Workflow-level weak links and acceptance-resource demand      | Observable resource loads and fully loaded cost per valid outcome           |
| Automation bias, learning, and deskilling [Bainbridge, 1983, Endsley and Kiris, 1995, Parasuraman and Manzey, 2010, Bastani et al., 2025, Lee et al., 2025] | Attention, situation awareness, reliance, and unaided learning    | Continuity obligation when authority and expertise move       | Acceptance labor and retained-capability debt as measurable operating terms |

option value, and dependency.

## 2.4 Limits of the research design

Public data systematically favor successes, measurable digital work, and organizations willing to disclose. Enterprise incidents, abandoned pilots, shadow use, and outsourced review labor have weak denominators. Many studies measure activity, short-horizon quality, or self-reported time rather than accepted economic value. Models and products change during observation windows. Company-level adoption may be endogenous to prior intangible investment, management quality, or worker composition. Recent working papers provide unusually relevant evidence but require replication.

The framework is also partly constructive. The specific runtime definition of *acceptance boundary*, along with *boundary load*, *acceptance-capacity runway*, *operational capacity shadow*, *authority surface*, *behavior bundle*, *AI human tax*, *protected recovery opportunity*, and *retained-capability debt*, are proposed analytical devices. The phrases are not claimed as coinages, standards, or validated scalar risk scores, and their parameters or forecasting performance have not been estimated across organizations. Equations below clarify measurable relationships; they do not claim a universal production function.

# 3 The boundary stack

## 3.1 From capability to consequence

The model frontier is not one line. Morris et al. [2024] separate performance depth, breadth or generality, and autonomy, while holistic evaluation shows that accuracy, calibration, robustness, fairness, toxicity, and efficiency do not collapse into one context-free ranking [Liang et al., 2023]. This paper

**Table 3:** Crosswalk from constructs to observables and propositions

| Construct                                                     | Analytical role                                                                                               | Illustrative observables                                                                                                      | Tests  |
|---------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|--------|
| Seven-boundary stack                                          | Antecedents and consequence constraints                                                                       | Capability threshold; validation error; credential scope; dependency map; recovery evidence; unit economics                   | P2–P6  |
| Operational capacity shadow and runway                        | Induced boundary-resource demand across workflows and time remaining before an SLO-constrained resource binds | Arrival and departure rates; service demand; allocation; utilization; work in process; tail lead time; provisioning lead time | P1, P4 |
| Behavior-bundle manifest                                      | Release, comparison, and evidence unit                                                                        | Resolved versions; effective permissions; tool trajectory; evaluator results; state diff                                      | P3, P5 |
| Four operating planes                                         | Proposed reusable shared-substrate functions with federated workflow ownership                                | Policy decisions; trace completeness; semantic SLOs; revocation and reconciliation drills; inherited-control evidence         | P3–P6  |
| Acceptance labor, protected recovery, and retained capability | Human cost, sustainable pace, and continuity                                                                  | Touch time; work intensity; after-hours load; unaided detection; calibration; takeover; apprenticeship progression            | P4, P7 |
| Venture and capital application                               | Delayed truth, selection, option value, and concentration                                                     | False negatives; novelty; rare outcomes; corroboration; provider/customer/financing concentration                             | P9     |

adds institutional boundaries that appear when performance is embedded in work.

The paper treats a **capability threshold** as an ex ante, workflow-local release criterion, not a benchmark label assigned after observing deployment results. For each versioned behavior bundle, owners declare a held-out task distribution and minimum business-quality rate  $q_w^{min}$ , repeated-run reliability  $r_{wk}^{min}$  over  $k$  independent trials, coverage and abstention requirements, and any latency or cost ceiling before comparing downstream configurations. The threshold is crossed only if the held-out estimates satisfy  $\widehat{\Pr}(Q_w = 1) \geq q_w^{min}$  and the prespecified repeated-reliability criterion  $\hat{r}_{wk} \geq r_{wk}^{min}$ , with uncertainty reported. Requirements may differ by consequence tier. Crossing this threshold establishes studied task capability; it does not establish authority, valid acceptance, or later value.

These boundaries are coupled. Better models can increase acceptance-ready probability and reduce verification work. They can also increase candidate arrival, action breadth, and dependency on inference capacity. Better observability can improve diagnosis but create privacy and retention risk. More human review can catch errors but overload experts or become ritual approval. The objective is therefore not to maximize any layer independently. It is to locate nondominated operating profiles within consequence-specific constraints.

### 3.2 Acceptance, valid acceptance, and realized value

For workflow  $w$ , let  $D_w = 1$  denote the observable decision to accept, commit, send, release, or otherwise make a candidate authoritative. Let contemporaneous validity be

$$Z_w = Q_w \wedge U_w \wedge P_w \wedge E_w \wedge S_w, \quad (1)$$

where  $Q$  is required business quality,  $U$  valid authority,  $P$  policy-conforming trajectory,  $E$  sufficient evidence, and  $S$  the required authoritative postcondition. A **validly accepted outcome** is  $A_w = D_w \wedge Z_w$ . For an internal draft,  $S$  may be a correctly stored artifact. For a payment, it includes permitted execution and reconciliation with the ledger. For an investment judgment, it means that the authorized decision and its evidentiary basis were recorded and approved; it does not mean the investment will later succeed. Let  $Y_w(h)$  record realized value or defect state at later horizon  $h$ .

This separation permits false acceptance,  $D = 1, Z = 0$ ; false rejection,  $D = 0, Z = 1$ ; and work that is validly accepted now but produces an adverse delayed outcome. It prevents three category

**Table 4:** Seven boundaries between frontier capability and organizational scale

| Boundary       | Question                                                                                                    | Typical failure                                                                                     | Minimum evidence                                                                                                        |
|----------------|-------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|
| Capability     | Can the configured system perform the task distribution at the required quality and reliability?            | Benchmark transfer failure; long-tail task failure; variance across runs                            | Local evaluation, pass-at- $k$ , tail cases, calibrated abstention                                                      |
| Epistemic      | Can the institution distinguish a valid result from a fluent error at acceptable cost?                      | Verification exceeds generation cost; sources are not authoritative; ground truth is delayed        | Provenance, independent checks, reviewer calibration, delayed-outcome audit                                             |
| Authority      | May this identity take this action for this purpose, now, within these limits?                              | Excess privilege, stale delegation, purpose drift, unowned autonomy                                 | Named principal and sponsor, scoped policy, expiry, limits, revocation test                                             |
| Systems        | Does the full path preserve invariants across data, memory, tools, services, and records?                   | Prompt injection, stale retrieval, tool misuse, partial state change, dependency drift              | End-to-end trajectory, data lineage, state diff, transaction and policy checks                                          |
| Organizational | Can downstream people and systems absorb, integrate, and decide on the arriving work at a sustainable pace? | Review queue, exception backlog, coordination overload, unsustainable work intensity, metric gaming | Queue and touch time, accepted throughput, exception and rework distributions, protected recovery and perceived control |
| Recovery       | Can authority be stopped and effects contained, reconciled, restored, and safely continued?                 | Kill switch stops calls but leaves harmful state; failover changes behavior                         | Timed drill, selective revocation, affected-state discovery, reconciliation, safe-continuity proof                      |
| Capital        | Do compute, provider, data, labor, and financing dependencies preserve viable unit economics?               | Token or cloud bill scales faster than value; provider concentration; valuation masks cash needs    | Fully loaded unit cost, concentration map, price/quota stress, portability test                                         |

errors. First, an API success is not necessarily a task success. Second, a correct final answer can be produced through an impermissible path. Third, a valid judgment can still have an uncertain delayed outcome. Immediate acceptance, contemporaneous validity, later defect escape, and realized business value must be measured separately.

Operational measurement requires a predeclared labeling protocol. Accountable domain and control owners define the  $Q$  and  $E$  rubrics, examples, tolerances, and adjudication horizon before a release;  $U$ ,  $P$ , and  $S$  should be derived from system records, policy decisions, and authoritative state checks wherever possible. Every label records the behavior-bundle version, decision time, label time, and applicable horizon. A blinded probability sample of accepted outcomes receives independent second review, with disagreements adjudicated separately and inter-rater reliability reported by consequence tier. Rejected candidates require sampled delayed adjudication as well: because most never receive the full downstream test, false rejection is otherwise unobservable. Audits should look for threshold gaming, selective escalation, and post hoc relaxation of validity definitions. These procedures do not make judgment objective; they make who judged what, against which rule, and when inspectable.

Consequence determines the proof burden. A useful consequence model assigns a tier or vector based on effects on money, rights, confidentiality, safety, external commitments, and systems of record.

It is assigned by accountable business and control owners before release. Overlapping effects inherit the most demanding applicable control. A generic “human in the loop” is not a tier: a fatigued click may be weaker than a deterministic transaction limit or independent state check.

### 3.3 Boundary load and accepted throughput

Let  $\mathcal{W}$  be the set of active workflows,  $\mathbf{a}_w$  workflow  $w$ ’s autonomy profile, and  $\boldsymbol{\alpha} = (\mathbf{a}_w)_{w \in \mathcal{W}}$  the organization’s joint profile. Let  $\lambda_w(\mathbf{a}_w)$  be workflow  $w$ ’s candidate-arrival rate. For a boundary-resource class  $j$ —for example automated validation, specialist review, security authorization, platform engineering, or incident response—let  $d_{wj}(\mathbf{a}_w)$  be expected service demand per candidate and  $B_j$  the organization’s shared service capacity per period. A first-order fluid load approximation is

$$\rho_j(\boldsymbol{\alpha}) = \frac{\sum_{w \in \mathcal{W}} \lambda_w(\mathbf{a}_w) d_{wj}(\mathbf{a}_w)}{B_j}, \quad \boldsymbol{\rho} = (\rho_j)_j, \quad \rho = \max_j \rho_j. \quad (2)$$

Under stable long-run averages, compatible units, a defined service discipline, and no abandonment, Little’s Law links work in process to throughput and lead time, while heavy-traffic approximations predict nonlinear delay as utilization and variability rise [Little, 1961, Kingman, 1961]. Under those diagnostic assumptions, sustained  $\rho \geq 1$  implies that unfinished work, rejection, or unperformed control must grow unless the organization adds capacity, reduces demand, reallocates priority, or changes the boundary. Shared specialists, security reviewers, platform teams, and incident responders make cross-workflow interference central rather than exceptional. Real acceptance systems also have priority classes, bursts, retries, rework loops, blocking dependencies, fatigue, and abandonment. Equation 2 is therefore a resource-centric fluid heuristic, not a fitted queueing model.

The model does not assume that all boundary work is human or fixed. Automated tests, policy engines, typed tools, schemas, transaction limits, and authoritative state checks can reduce  $d_{wj}$  and increase  $B_j$ ; process redesign can remove obsolete gates; a stronger model can reduce or increase review demand. Let  $b_{wj}$  be the effective share of resource  $j$  made available to workflow  $w$  under the declared service discipline and priority policy, let  $\mathbf{b} = (b_{wj})$ , require  $\sum_w b_{wj} \leq B_j$ , and let  $\mathcal{J}_w$  contain the resources required by that workflow. Let  $p_w(\mathbf{a}_w, \boldsymbol{\rho})$  be the probability that a processed candidate is validly accepted, allowing congestion to affect attention, abandonment, and error. A simplified valid-acceptance-throughput approximation is

$$T_w^{VA}(\boldsymbol{\alpha}, \mathbf{b}) = p_w(\mathbf{a}_w, \boldsymbol{\rho}) \min \left\{ \lambda_w(\mathbf{a}_w), \min_{j \in \mathcal{J}_w} \frac{b_{wj}}{d_{wj}(\mathbf{a}_w)} \right\}. \quad (3)$$

The equation exposes a scale asymmetry. If autonomy doubles  $\lambda$  while shared capacity, service demand, and allocation remain unchanged, candidate output can double while validly accepted throughput remains capacity-bound. If stronger models or controls increase  $p$ , lower  $d_j$ , or remove a handoff, the same autonomy increase can create durable value. More autonomy can also reduce context reconstruction across handoffs. The weak-link analogy to Amdahl’s Law is intuition, not causal proof about any observed deployment [Amdahl, 1967].

The economically relevant denominator is the validly accepted outcome. For measurement horizon  $h$ ,

$$C_w^{VA}(h) = \frac{C_w^{\text{runtime}}(h) + C_w^{\text{data}}(h) + C_w^{\text{assure}}(h) + C_w^{\text{labor}}(h) + C_w^{\text{incident}}(h) + C_w^{\text{exit}}(h)}{N_w^{VA}(h)}. \quad (4)$$

This prevents token cost, license cost, or labor time from masquerading as full economics. The numerator includes data acquisition and indexing, evaluation and observability, compliance, support,

opportunity cost, direct incident response and remediation, and provider exit where material. The denominator should be stratified by quality and consequence so that cheap false acceptance is not counted as productivity. The formulation also makes an important counterargument testable: if stronger models remove enough verification and recovery work,  $C^{VA}$  should fall even at higher nominal inference cost.

### 3.4 Acceptance-capacity runway: scaling before the boundary binds

The **operational capacity shadow** is induced demand on data stewardship and serving, identity, security, evaluation, review, telemetry, technical recovery, financial control, and human acceptance work. It grows as candidate production or delegated authority expands, and the load model can make it prospective. Only recurrent service demand with compatible units enters  $\rho_j$ . Data integrity, legal authority, maximum credible harm, and protected recovery opportunity may instead remain release gates or feasibility constraints; human review or incident-response time enters the load model only when it is defined as a service resource. For interpretation, decompose  $\lambda_w = x_w y_w$ , where  $x_w$  is root workflow requests per period and  $y_w$  is candidates generated per root request. Likewise, write  $d_{wj} = m_w v_{wj} s_{wj}$ , where  $m_w$  is executed trajectories per candidate, including delegation and retries,  $v_{wj}$  is expected visits to resource  $j$  per trajectory, and  $s_{wj}$  is service demand per visit. Peak concurrent agents are not multiplied into mean load when their trajectories are already counted in  $m_w$ ; concurrency instead informs arrival variability, correlated failure, and tail delay.

For scenario  $\omega$ , information set  $\mathcal{I}_t$ , and upper predictive quantile  $1 - \gamma$ , define

$$\tilde{\rho}_{j,\gamma}^\omega(t + \Delta) = \text{Quantile}_{1-\gamma} \left[ \frac{\sum_{w \in \mathcal{W}} \lambda_w^\omega(t + \Delta) d_{wj}^\omega(t + \Delta)}{B_j^\omega(t + \Delta)} \middle| \mathcal{I}_t \right], \quad (5)$$

where future  $B_j$  includes only capacity already committed and forecast to be recruited, procured, automated, tested, and released by that horizon. Uncertainty in demand, service demand, and committed-capacity readiness belongs inside the predictive distribution. Let  $\rho_j^{\text{SLO}} < 1$  be the highest utilization consistent with the resource's consequence-specific tail-latency, error, abandonment, and recovery objectives. No universal 70 or 80 percent threshold is assumed. Define the first forecast crossing and remaining runway as

$$\begin{aligned} \tau_{j,\gamma}^\omega(t) &= \inf \left\{ \Delta \geq 0 : \tilde{\rho}_{j,\gamma}^\omega(t + \Delta) \geq \rho_j^{\text{SLO}} \right\}, \\ \mathcal{R}_{j,\gamma}^\omega(t) &= \tau_{j,\gamma}^\omega(t) - L_j^{\text{plan}}, \quad \mathcal{R}_\gamma^\omega(t) = \min_j \mathcal{R}_{j,\gamma}^\omega(t), \end{aligned} \quad (6)$$

where  $0 < \gamma < 1$  is the forecast-tail probability,  $L_j^{\text{plan}}$  is a conservative estimate of the lead time to make additional capacity safely available, and  $\mathcal{R}_\gamma^\omega$  is the organization's shortest resource runway. The infimum of an empty crossing set is  $\infty$ . When  $\mathcal{R}_{j,\gamma}^\omega \leq 0$ , the organization is late under that scenario unless it shapes arrivals, reduces service demand, reallocates capacity, or changes the authority profile. Hard exposure boundaries—for example privileged mutation without selective revocation or regulated data without an enforceable purpose rule—remain release gates regardless of forecast utilization.

Runway is a proposed decision variable, not an estimated empirical law. Scenarios should be reported separately when defensible probabilities do not exist, and forecasts should be preserved at decision time so that interventions do not make the prediction appear falsely wrong. The construct is weakened if shorter forecast runway does not predict earlier resource-SLO breach better than simpler baselines based on seats, tokens, or lagged utilization.

### 3.5 The autonomy sweet spot

The **autonomy sweet spot** is not “low-risk tasks” or “copilot before autopilot.” Autonomy is a profile rather than a scalar:  $\mathbf{a} = (a_{goal}, a_{plan}, a_{data}, a_{tool}, a_{act}, a_{external}, a_{persist}, a_{delegate})$ , covering goal choice, planning, data access, tool selection, state-changing action, external communication, persistence, and delegation. Because workflows share resources, the decision unit is the joint profile and allocation, abbreviated  $\mathbf{x} = (\boldsymbol{\alpha}, \mathbf{b})$ . For common horizon  $h$ , define

$$\mathcal{F}_{\mathcal{W}}(h) = \left\{ (\boldsymbol{\alpha}, \mathbf{b}) : \begin{array}{l} \mathbf{a}_w \in \mathcal{A}_w, \quad \sum_w b_{wj} \leq B_j, \\ g_{kw}(\boldsymbol{\alpha}, \mathbf{b}; h) \leq 0 \quad \forall k, w \end{array} \right\} \quad (7)$$

where the constraints include thresholds for defect escape, policy conformance, latency, recovery time, privacy, concentration, retained human capability, tail loss, maximum credible harm, and spillovers onto other workflows’ service levels. The organizational sweet spot is the nondominated region

$$\begin{aligned} \mathcal{S}_{\mathcal{W}}(h) = \text{Pareto}_{\mathbf{x} \in \mathcal{F}_{\mathcal{W}}(h)} \left( \sum_w V_w N_w^{VA}(\mathbf{x}; h), - \sum_w C_w(\mathbf{x}; h), \right. \\ \left. - \sum_w \mathbb{E} L_w^{escape}(\mathbf{x}; h), - \sum_w R_w^{cap}(\mathbf{x}; h) \right). \end{aligned} \quad (8)$$

Here  $V_w$  is value per validly accepted outcome;  $N_w^{VA}$  is the expected valid-outcome count, approximately  $hT_w^{VA}$  under stable within-horizon rates;  $C_w$  is fully loaded operating cost;  $L_w^{escape}$  is uncompensated downstream harm and externality from escaped failures; and  $R_w^{cap}$  is retained-capability risk, all over the same horizon. Direct incident-handling and remediation spend belongs in  $C_w$  and is excluded from  $L_w^{escape}$  to avoid double counting. Rights, fiduciary duties, safety rules, and legal prohibitions remain hard constraints even when modeled expected value is positive. A single joint profile exists only after accountable decision-makers supply a permissible utility function, weights, or dominance rule; otherwise multiple profiles can be incomparable. A workflow-local sweet spot is a conditional slice that holds  $\boldsymbol{\alpha}_{-w}$  and  $\mathbf{b}$  fixed and preserves spillover constraints. These quantities may be hard to estimate, but leaving them implicit does not make them disappear.

Work moves toward higher autonomy when outcomes are frequent, bounded, observable, cheaply verifiable, reversible or compensable, and linked to authoritative data. It moves toward constrained recommendation when objectives are ambiguous, feedback is delayed, actions are irreversible, externalities are high, or verification requires the same scarce judgment as production. Controls can move the sweet spot: a high-consequence payment workflow can support more automation with hard transaction limits, separation of duties, and deterministic reconciliation than a low-dollar but reputation-sensitive communication whose correctness is difficult to formalize.

### 3.6 The behavior bundle as the release unit

The base model is too narrow a unit for production reliability. Behavior can change through system instructions, routing, orchestration, tool schemas, retrieval indexes, memory policy, identities, permissions, validators, budgets, or fallbacks. This paper proposes a **behavior-bundle manifest**

$$\mathcal{B} = \{\text{model, runtime, knowledge, tools, identity,} \\ \text{policy, evidence, release, recovery}\}. \quad (9)$$

The **model** field records route and sampling configuration; **runtime** records application, orchestrator, package, and image artifacts; **knowledge** records retrieval, memory, source, index, freshness, lineage,

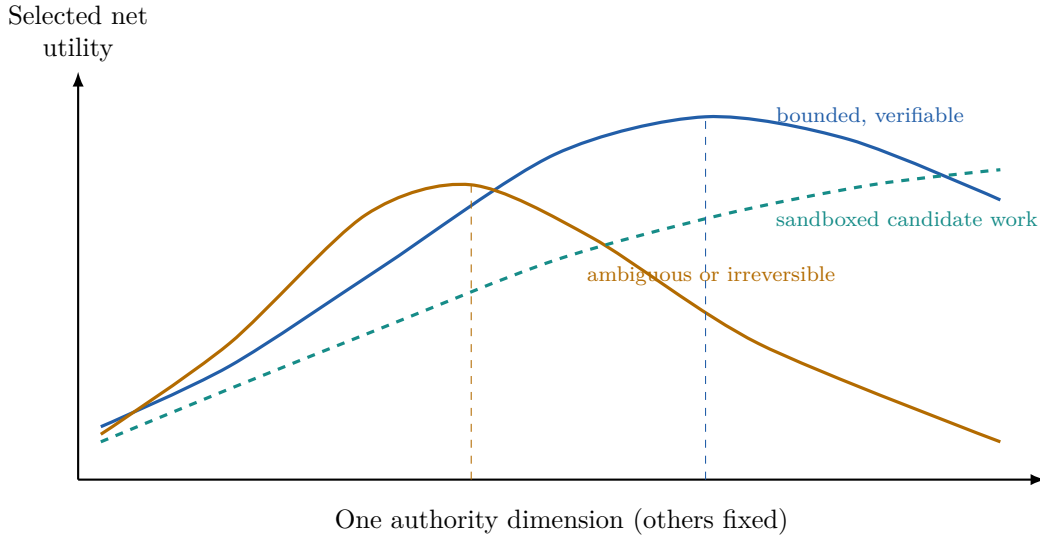


**Figure 2:** Illustrative one-dimensional slices through an autonomy profile after selecting a permissible scalar utility and holding other profiles and allocations fixed. Different workflows can have different shapes, discontinuities, or no interior maximum. The curves are hypotheses, not estimated functions.

tenant, and access-control state; **tools** records typed contracts, protocol versions, transactions, and compensation logic; **identity** records principals, delegation, credentials, network, and egress; **policy** records budgets, retries, loops, and delegation limits; **evidence** records evaluation, telemetry, redaction, retention, and audit integrity; **release** records canary and promotion policy; and **recovery** records fallback, reconciliation, degraded mode, and tested restoration. A release records immutable artifacts or controlled references. Each execution resolves them into evidence containing the actual route, versions, effective permissions, tool results, policy decisions, validators, and authoritative state changes.

The goal is a better evidence trail and tested restoration, not identical reproduction of stochastic text or guaranteed causal reconstruction. Hidden provider changes, dynamic data, and external-service state can prevent exact replay. A lineage graph and continuous evaluation complement the manifest in dynamic environments. The proposal is weakened if model identity alone predicts production behavior across relevant changes, or if manifest versioning adds cost without reducing change failures, investigation time, or recovery uncertainty.

The empirical question is whether observed gains attenuate at the validation, integration, authority, or human-load transitions represented by these constructs.

## 4 Evidence of gains, attenuation, and shifted work

### 4.1 Local productivity is real and heterogeneous

The strongest evidence rejects both blanket optimism and blanket skepticism. [Brynjolfsson et al. \[2025\]](#) studied a deployed generative-AI assistant in customer support and found an average productivity gain concentrated among less experienced and lower-skilled workers, alongside evidence consistent with faster learning. [Cui et al. \[2026\]](#) pooled three randomized field experiments; because compliance was imperfect, its preferred instrumental-variables estimate is a local average treatment effect for developers induced by assignment to use the tool: 26.08 percent more completed software tasks, with a standard error of 10.3 percent. The individual site estimates were noisy, and use and

gains were larger for less experienced developers. Dell’Acqua et al. [2026b] found that consultants using GPT-4 completed 12.2 percent more tasks, worked 25.1 percent faster, and produced higher-quality work on tested tasks inside the model frontier, while participants were 19 percent less likely to solve a selected outside-frontier task correctly.

These findings establish causal benefits under studied configurations. They do not identify one enterprise-wide coefficient. The support setting had structured outcomes and a rich organizational knowledge base; the coding trials measured completed tasks under specific tools; the consulting experiment used controlled tasks and a dated model. Worker experience, task location relative to the frontier, adoption intensity, and complement quality all matter.

The negative evidence is equally conditional. In a randomized preprint involving 16 experienced open-source developers working on 246 issues in familiar mature repositories, early-2025 AI tools made participants 19 percent slower even though they believed they were faster [Becker et al., 2025]. That result should not be generalized to all developers or later models. It demonstrates that perceived acceleration and end-to-end time can diverge where context acquisition, verification, and mature-code constraints are substantial.

## 4.2 The production hierarchy attenuates generation gains

The clearest evidence for a boundary mechanism comes from software, where intermediate states are observable. Demiret et al. [2026] link usage telemetry to more than 100,000 GitHub developers. Estimated cumulative effects increased across autocomplete, interactive agents, and autonomous agents – roughly 40, 140, and 180 percent for commits – but the largest effect attenuated to about 50 percent for projects and 30 percent for releases. Across four app marketplaces, the authors found a moderate increase in new apps but no increase in total usage. The working paper interprets this through a weak-link production function with strong complementarity between AI and human effort. Its matched event-study design is not a randomized organization-wide experiment, and commits, projects, releases, and usage each capture different margins. The attenuation is nevertheless directly aligned with Equation 3.

The longitudinal enterprise study by He et al. [2026] observes the same mechanism inside one firm. Across 802 developers and 196,212 pull requests from January 2024 through April 2026, merged pull requests per capita reached 2.09 times the pre-mandate baseline. Under stricter calendar controls, the within-developer estimate associated with adoption and accumulated use was about 1.46 times. The organization was young, AI-forward, and offered uncapped access, making it closer to an upper-envelope case than an average firm. Adoption timing and intensity were not randomized.

The downstream change is the crucial result. Raw pull-request volume rose faster than the reviewer pool, per-reviewer load roughly doubled, automated review coverage rose above human coverage, the share of pull requests with human review fell, and substantive human comments became less common. AI-authored changes took longer in human review even after observable controls. Merge and revert rates did not collapse, but those are coarse short-horizon quality measures that miss maintainability, security, and incidents. The case shows adaptation, not simple failure: the firm preserved flow by automating more review. It also shows why merged pull requests alone are incomplete governance.

Google’s deployed code-review work illustrates the favorable version of the same logic. A narrow system that proposes edits in response to reviewer comments addresses 7.5 percent of reviewer comments through an applied ML suggestion and was estimated to save hundreds of thousands of engineer-hours annually at Google scale [Frommgen et al., 2024]. The task begins with a human-identified issue, has local context, produces a proposed patch, and remains inside an established review and merge path. It is not “general coding automation.” It is a high-volume, bounded, acceptance-rich workflow.

**Table 5:** Selected empirical results and the boundary they expose

| Study and status                          | Design                                                                   | Result used                                                                                           | Boundary interpretation and limit                                                                             |
|-------------------------------------------|--------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|
| Cui et al. [2026], peer reviewed          | Three randomized field experiments; 4,867 developers                     | 26.08% pooled IV/LATE for use among developers induced by assignment; heterogeneous sites             | Strong task-level causal evidence for compliers; not a measure of releases, incidents, or firm value          |
| Dell’Acqua et al. [2026b], peer reviewed  | Preregistered experiment; 758 consultants                                | Faster, more, and higher-quality work inside frontier; 19% lower correctness on selected outside task | Task routing and verification are part of value; model and task set are dated and bounded                     |
| Dell’Acqua et al. [2026a], peer reviewed  | Preregistered field experiment; 791 P&G professionals                    | Individuals with AI matched teams without AI; human judgment retained value in selection              | Generation and evaluation have different comparative advantages; innovation exercise is not production launch |
| Demirer et al. [2026], NBER working paper | Telemetry and matched event study; >100,000 GitHub developers            | Up to 180% commits, 50% projects, 30% releases; no app-usage increase                                 | Direct production-hierarchy attenuation; non-random usage and digital-work scope                              |
| He et al. [2026], preprint                | Single-firm panel; 802 developers, 196,212 PRs                           | 2.09x per-capita PRs; reviewer load about doubled; review shifted to automation                       | Near-best-case throughput with downstream adaptation; no random rollout and coarse quality proxies            |
| Becker et al. [2025], preprint            | Randomized trial; 16 experienced developers, 246 mature-repository tasks | 19% slower with AI while expecting to be faster                                                       | Perceived speed can miss context and verification costs; small, specialized sample                            |
| Cowgill et al. [2026], peer reviewed      | Hiring and startup-investing experiments                                 | AI access lowered screening accuracy 4–9% on average and raised demand for other signals 6–9%         | Candidate polish can increase evaluator work; some non-English-speaking senders became easier to assess       |

### 4.3 The AI human tax, protected recovery, and retained-capability debt

The phrase **AI human tax** denotes human work required to turn machine output into accepted outcomes. It includes framing, context preparation, verification, correction, exception handling, integration, governance, incident response, recovery, and maintenance of unaided capability. The term does not imply that all such work is waste. Four components should be separated:

1. **temporary adoption cost:** learning, workflow redesign, data preparation, and tool migration;
2. **avoidable rework:** hallucination cleanup, duplicate candidates, poor routing, brittle integrations, and low-value approval rituals;
3. **inherited bureaucracy:** gates and coordination costs that predate AI and may be exposed rather than caused by it; and
4. **accountability and resilience work:** consequence-proportionate proof, exception ownership, reconciliation, and safe continuity that cannot be wished away by faster generation.

The productivity J-curve predicts that general-purpose technologies require complementary intangible investment and can show delayed measured returns [Brynjolfsson et al., 2021]. That makes early human tax a cohort variable, not a permanent constant. A useful deployment should reduce avoidable work over time while preserving necessary proof. If human minutes per accepted outcome remain flat or rise after maturation, the organization should test whether model quality, workflow design, data, or governance is the cause.

Candidate inflation can also impose a cost on people who did not choose the tool. An editorial analysis of one management journal reports a 42 percent rise in submissions after ChatGPT’s release alongside lower measured writing quality and less diverse AI-mediated reviews [Gartenberg et al., 2026]. Detection and causal attribution are contestable, and one journal is not the economy. The case illustrates a review-queue externality: cheap production can transfer screening cost downstream.

That transferred load has a pace and recovery dimension. **Protected recovery opportunity** is the organization-provided time, staffing, control, and relief from demand that allow workers and teams to restore attention and judgment between periods of demand. It is not a diagnosis, a personal trait, or an invitation to treat people as another autoscaling pool. In the 2022 OECD surveys of 5,334 workers and 2,053 firms in finance and manufacturing across seven countries, workers were generally positive about AI’s effects, but roughly three quarters of AI users also reported that it increased their work pace; the survey did not establish whether the pace was excessive or causally produced distress [Lane et al., 2023]. A working paper using 2004–2023 U.S. time-diary data associates higher occupational AI exposure with longer workdays and less leisure, but exposure measures, selection, and observational design prevent a causal claim about generative AI or burnout [Jiang et al., 2025]. An ILO working paper identifies surveillance, work intensification, reduced autonomy, privacy, and data-driven management as plausible psychosocial risks; it is a legal and policy synthesis, not a prevalence estimate [Karimova, 2026].

The resulting responsibility is organizational. WHO guidance prioritizes organizational interventions that change working conditions, alongside manager training, worker support, accommodation, and return-to-work measures [World Health Organization, 2022]. For AI scaling, that means shaping arrival rates, staffing review and incident rotations, limiting concurrent trajectories and after-hours escalation, preserving decision latitude, protecting recovery time, and involving workers or their representatives in workflow design. A break reminder, check-in, or “recovery prompt” may be a voluntary aid, but it does not compensate for an unbounded queue, an impossible service objective, or accountability without authority. Such prompts should be non-diagnostic, privacy-preserving, and never used as covert performance or health surveillance.

Human-recovery indicators belong beside the boundary-load model: sustained task concurrency, interruptions and alerts, after-hours work, queue age, incident-call load, protected recovery time, perceived control, psychological safety, turnover, absence, and error under fatigue. Because several measures are sensitive, the default should be aggregate cohort analysis, data minimization, restricted access, and worker participation rather than individual scoring. Platform owners can supply rate limits, workload telemetry, quiet periods, rotation support, and escalation controls; domain leaders and People or occupational-safety functions remain accountable for job design and support. The shared operating substrate can make overload visible and constrain machine demand, but it cannot automate away the employer’s responsibility for sustainable work.

**Retained-capability debt** is the accumulated future cost and operational risk created when an organization consumes practice, apprenticeship, or independent judgment faster than it replenishes the human capability needed to specify, verify, challenge, and recover from AI-mediated work. As a conceptual accounting identity rather than an estimated law, its stock can be written  $RCD_{t+1} = \max\{0, RCD_t + q_t - r_t\}$ , where  $q_t$  is required practice or learning displaced during the period and  $r_t$  is replenishment through unaided practice, apprenticeship, simulation, and recovery drills. It is the paper’s narrower operational substitute for the colloquial phrase “AI brain rot”; neither *capability debt* nor the underlying deskilling mechanism is claimed as new. No retained source establishes clinical neurological damage from enterprise AI use. The evidence concerns effort, learning, calibration, and automation bias. A CHI survey found self-reported reductions in critical-thinking effort associated with confidence in generative AI, but its cross-sectional self-report design cannot establish skill loss [Lee et al., 2025]. A PNAS experiment found that unguarded generative-AI assistance could improve practice performance while harming later unaided learning; a purpose-built tutor with guardrails mitigated the effect [Bastani et al., 2025]. Classic automation research shows why monitoring a reliable system can weaken attention, reduce situation awareness, and leave humans handling only unusual failures [Bainbridge, 1983, Endsley and Kiris, 1995, Parasuraman and Manzey, 2010, Lyell and Coiera, 2017]. A 2025 multicenter observational study reported lower unaided adenoma detection

after exposure to AI-assisted colonoscopy; its before/after design cannot rule out workload, case-mix, time, or behavioral confounding and should be treated as a prompt for prospective measurement, not a general workplace effect [Budzyń et al., 2025].

Retained-capability debt is therefore measured through performance, not rhetoric: unaided error detection, reviewer calibration, exception resolution, recovery drill success, junior-to-senior progression, and time to safe manual or degraded operation. The hypothesis is weakened if assisted organizations preserve or improve these measures without explicit practice, or if automated recovery consistently outperforms retained human capability across relevant failures.

These workflow mechanisms move the unit of analysis to the organization: who supplies shared acceptance capacity, who owns local validity, and whether workload, recovery opportunity, and future reviewer capability scale with candidate volume.

## 5 Organizational evidence and the emerging sweet spot

### 5.1 Adoption is not operating scale

Access, use, accepted outcomes, and financial value are separate stages. The 2026 AI Index summarizes survey evidence that AI use has become widespread across organizations while scaled agent deployments remain much less common [Stanford Institute for Human-Centered Artificial Intelligence, 2026]. National statistics show the same definitional problem. U.S. Census and Eurostat estimates measure whether an organization used at least one qualifying AI technology, not whether a consequential workflow produced net value [Grundy et al., 2026, Eurostat, 2025]. These adoption denominators are important, but they do not measure Equation 1.

A large 2026 preprint provides unusually detailed organizational usage evidence but uses three distinct samples. In an aggregate ChatGPT Enterprise organization-week panel whose organization count is not disclosed, output tokens rose approximately sevenfold from June 2025 to March 2026 and roughly fourfold among organizations that had adopted by June 2025. A separate worker-characteristics sample contained 1,764 organizations and 17,446,551 ChatGPT Enterprise messages. In a separate public-company comparison, larger and more software-, R&D-, and SG&A-intensive firms adopted more often, while larger adopters used fewer messages per employee [Chatterji et al., 2026]. The authors are OpenAI employees or paid contractors, and the telemetry excludes personal accounts, API use, and competing products. The design establishes usage patterns and organizational correlates, not causation or return. The interpretation consistent with the boundary model is that larger firms can fund shared complements but must diffuse them through more heterogeneous permissions, processes, and teams.

Interviews in a RAND study supply a useful failure taxonomy—weak business problem, proxy mismatch, data quality, inadequate infrastructure, leadership expectations, and treating deployment as an endpoint—but do not estimate an AI-project failure rate because they elicited perceived mechanisms from 50 selected industry practitioners rather than sampling a defined portfolio of projects [Ryseff et al., 2024]. This distinction is symptomatic of the paper’s method: mechanism evidence is retained without converting it into a prevalence statistic.

The cross-government Microsoft 365 Copilot experiment makes those complements visible. Twenty thousand licenses were distributed across 12 U.K. public organizations during a three-month uncontrolled trial; telemetry covered about 14,500 users, and 7,115 survey responses were received. Adoption reached about 80 percent among telemetry-covered users under a lenient 30-day definition, and survey respondents reported saving 26 minutes per day on average. The same report documents uneven feature knowledge, training and support needs, permission and content-access problems, difficulty

with nuance, and continuing oversight [Government Digital Service, 2025]. Telemetry was unavailable for roughly 5,500 licensees, response selection was not random, and reported time was not audited or traced to redeployment. The stronger inference is that one product crossed twelve different operating environments and encountered organizational boundary work even when access was centrally procured.

## 5.2 Organization design moves with the production function

AI-native startups provide a different view because they are less constrained by installed roles and systems. A working paper comparing Y Combinator batches W20–F24 with U.S. PitchBook startups first financed in 2020–2024 estimates that AI-labeled YC firms were about 25 percent smaller and classifier-identified AI firms in the broader sample about 12 percent smaller. Both samples show a more engineering-intensive composition and fewer entry-level employees and managers. The YC valuation difference was statistically insignificant, while the broader and selectively observed valuation sample showed higher valuations for AI firms [Kim and Koning, 2026]. The estimates are cross-sectional; labels, classifier error, employment coverage, missing valuations, young-firm survival, and a short window limit causal interpretation. Still, the pattern implies that AI may change the occupational composition of growth rather than simply remove a constant fraction of every job.

That composition creates a second-order constraint. Flatter firms can generate and iterate with fewer handoffs, but they may also narrow apprenticeship, concentrate review in senior staff, and increase dependence on encoded operating knowledge. The relevant question is therefore not headcount alone. It is whether accepted throughput, protected recovery opportunity, and the production of future reviewers scale with the firm. This is why retained-capability measures belong beside assisted output.

## 5.3 Named deployments locate, but do not prove, the sweet spot

Table 6 compares deployments with enough detail to locate the automated stage and retained boundary. The cases are not a league table and do not estimate prevalence.

Three regularities recur. First, the stronger cases begin with a named workflow and owner, not a general license mandate. Second, acceptance evidence is legible: a resolved support case, an applied code edit, a published course unit, or an advisor-approved communication. Third, higher-volume automation retains an escape path through edit, exception routing, deterministic checks, or human service. This does not prove those design features caused the reported results. It identifies conditions that can be tested across cases.

## 5.4 The sweet spot is a boundary surface

The evidence does not support one autonomy level for a profession or company. A deployment is indexed by workflow  $w$ , consequence  $c$ , automation stage  $g$ , behavior bundle  $\mathcal{B}$ , and environment  $e$ . The sweet spot is therefore a feasible region over  $(w, c, g, \mathcal{B}, e)$ , not a point on a universal curve. Stages matter: an AI system may autonomously acquire and normalize data, recommend during analysis, require human selection, and execute only through deterministic limits. This stage-specific view follows longstanding automation research rather than treating “human in the loop” as one setting [Parasuraman et al., 2000].

Table 7 states the comparative prediction. These are moderators, not a scoring standard.

The practical sweet spot is thus neither permanent augmentation nor maximal autonomy. It is the set of nondominated authority profiles that satisfy the applicable consequence, shared-capacity,

**Table 6:** Named deployments as boundary observations

| Case; source type                 | Observed production use                                                                                                                                                              | Acceptance boundary                                                                                            | Inference limit                                                                                                                                                                        |
|-----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Customer support;<br>CAUSAL STUDY | 5,172 agents; 15.2% more issues resolved per hour in the preferred estimate, with larger gains for less-experienced workers [Brynjolfsson et al., 2025]                              | Suggestions could be accepted, edited, or ignored; resolution and customer measures supplied frequent feedback | One firm and mature workflow; not full autonomy or enterprise profit                                                                                                                   |
| Morgan Stanley; DEPLOYMENT REPORT | Vendor reports more than 98% of advisor teams actively using an assistant grounded in a roughly 100,000-document corpus [OpenAI, 2025]                                               | Controlled knowledge retrieval and regression evaluation; advisors remain responsible for client-facing work   | Vendor-authored; no controlled revenue, quality, or compliance outcome                                                                                                                 |
| BBVA; DEPLOYMENT REPORT           | Bank expanded to 11,000 licenses; reporting users said they saved 2.8 hours per week on average, with respondent count undisclosed [BBVA, 2025]                                      | Legal, compliance, security, training, and internal communities accompanied access                             | Self-report; redeployment and accepted value are not established                                                                                                                       |
| Klarna; AUTHORITATIVE RECORD      | Filing reports 31 million cumulative conversations since launch, 80% of 2025 service chats, separately dated handling and repeat measures, and 2024 savings [Klarna Group plc, 2026] | Dual-track service preserves a human option; no generative AI in credit underwriting                           | Mixed periods and internal methods; SEC correspondence required fuller disclosure of quality concerns [U.S. Securities and Exchange Commission, Division of Corporation Finance, 2025] |
| Duolingo; MIXED RECORD            | Company announced 148 courses in under one year; later filing reported a large increase in published course units [Duolingo, Inc., 2025, 2026]                                       | Shared content and AI increase production; human learning experts validate release                             | AI contribution is confounded with tooling; learning efficacy and review effort are not disclosed                                                                                      |
| U.K. government; GOVERNMENT STUDY | 20,000 licenses across 12 organizations; broad but heterogeneous use [Government Digital Service, 2025]                                                                              | Users retain responsibility; data access, training, support, and nuance define local boundaries                | Self-reported time and selected focus groups; no audited organization-level return                                                                                                     |

spillover, and recovery constraints. Better models, data, tools, controls, and feedback can expand that feasible region. Increased candidate volume, broader credentials, or weaker review can contract it even when benchmark capability improves. If accountable decision-makers select a permissible scalar utility, one profile may then be preferred; the cases in this review do not identify that optimum.

The organizational design question is therefore how to make common controls reusable without centralizing workflow judgment. Section 6 proposes one such allocation.

## 6 Where IT meets the pavement

The operational boundary is not a prompt template. It is a state-integrity problem spanning distributed systems, identity, data governance, security, service management, human factors, and finance. The proposed reference architecture specifies minimum control functions through four planes and a logically shared cross-plane data responsibility; it has not been empirically validated as a universal minimum. The planes are logical responsibilities; a managed platform may implement several of them, and building all components internally is not implied. The shared operating substrate is logically central but need not be one team, product, host, or failure domain.

**Table 7:** Conditions predicted to shift the feasible autonomy region

| Dimension                     | Toward more bounded autonomy                                                                     | Toward recommendation or manual control                                                           |
|-------------------------------|--------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| Objective and feedback        | Measurable postcondition; fast, authoritative feedback                                           | Ambiguous objective; delayed or contested ground truth                                            |
| Verification                  | Independent, cheaper than production; high error visibility                                      | Same scarce expertise as production; plausible errors are hard to detect                          |
| Consequence and recovery      | Bounded loss; reversible or compensable action; tested reconciliation                            | Irreversible effects; rights, safety, or reputation exposure; weak recovery                       |
| Task structure                | Decomposable, repetitive, typed interfaces, low sequential dependence                            | Tacit context, novelty, long dependencies, correlated error                                       |
| Authority                     | Narrow tools, least privilege, transaction and rate limits                                       | Broad credentials, delegation chains, unclear sponsor or expiry                                   |
| Boundary capacity             | Automated checks and spare review capacity; stable exception rate                                | Saturated reviewers, growing work in process, rising defect escape                                |
| Human continuity and recovery | Periodic unassisted practice, tested takeover, protected recovery, and manageable exception load | Automation removes practice or intensifies work while leaving humans responsible for rare failure |

## 6.1 Shared operating substrate: a platform, not a committee

AI does not make information technology less central. Once models can invoke tools and alter records, each shared service needs a named owner, although implementation may be federated across IT, Data, Platform Engineering, Security, managed providers, and jointly governed services. The labels matter less than the collective mandate: model and provider mediation, governed data access, agent identity, tool registration, policy enforcement, evaluation hooks, trace integrity, cost attribution, emergency revocation, and recovery evidence. One executive sponsor is accountable for cross-plane integrity, while shared-service owners operate their mechanisms and accountable business executives accept workflow-specific residual exposure.

Central control does not mean centralized business judgment. Workflow owners remain accountable for intended outcomes, authoritative postconditions, exception semantics, and accepted results; data and state owners define permissible use and mutation; security and reliability owners set control and incident baselines; finance governs value and concentration; People or occupational-safety functions govern sustainable job design; and an executive risk owner accepts residual exposure. Consequential workflows also require assurance sufficiently independent of both the producing domain and the platform. The proposed organizational form is therefore a reusable central spine with federated domain ownership and independent challenge, not a universal approval committee.

This configuration is a hypothesis, not a proven optimum. Logical centrality can remove duplicated entitlements and make limits, evidence, and selective revocation consistent. It can also create a queue, an opaque policy monopoly, a surveillance surface, or a correlated failure domain. Each inherited control therefore needs a named owner, service objective, evidence source, tenant and environment isolation, change policy, and tested degraded mode. Logical sharing must not imply one administrative, runtime, regional, or evidence failure domain: selective revocation, last-known-good policy, and containment evidence must remain usable when the primary platform is unavailable or untrusted. Product teams should be able to identify which controls they inherit, which remain local, and how both layers recover.

**Table 8:** Four-plane architecture for consequential AI workflows

| Plane              | Primary responsibility                                                                                                     | Minimum controls                                                                                                                                                                                                                                                                | Evidence of readiness                                                                                                                                                                                |
|--------------------|----------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Workload           | Execute a bounded task and preserve business-state invariants                                                              | Governed ingress; explicit state machine; model/router; retrieval and memory; typed tool broker; pre/postconditions; idempotency; backpressure; retry, loop, and delegation limits                                                                                              | Local and end-to-end task tests; state diff; load and exception behavior; isolated development, staging, and production                                                                              |
| Control            | Evaluate, distribute, and evidence policy; the authoritative-state owner independently enforces final action authorization | Inventory, owner, consequence tier, distinct agent identity, on-behalf-of chain, audience and purpose binding, delegation attenuation, least privilege, separation of duties, short-lived credentials, spend and transaction limits, revocation                                 | Policy decision log; entitlement review; negative authorization tests; expiry, selective revocation, control-service outage and compromise, policy rollback, and independent-audit-continuity drills |
| Evidence           | Determine what happened, whether it conformed, and what it cost                                                            | End-to-end correlation; versioned manifest; retrieval/tool provenance; policy and evaluator results; semantic service-level objectives; redaction, retention, access, and audit integrity                                                                                       | Valid-acceptance rate; false acceptance/rejection; defect escape; cost; human touch time; trace completeness; incident-derived regression                                                            |
| Technical recovery | Stop authority, discover affected state, reconcile, restore, and continue safely                                           | Kill/pause; identity revoke; tool/provider isolation; data and memory quarantine; checkpoints; compensation; known-good bundle; degraded mode; manual mode only where staffing, authority, current procedures, and unaided performance are load-tested; behavioral re-open gate | Timed containment and state-discovery drill; recovery time and point objectives; reconciled records; passed behavioral validation                                                                    |

The data responsibility crosses all four planes: source authorization, lineage, freshness, tenant access-control propagation, index build, poisoning quarantine, retention and deletion, residency, and citation-to-source validation. Retrieval-augmented generation cannot create authority that the underlying content lacks. It often reveals duplicate documents, stale ownership, accumulated permissions, and tacit knowledge that was never made operationally accessible. Large-scale audits of dataset licensing and attribution show that provenance and use rights are not reliably inferable from model availability [Longpre et al., 2024].

## 6.2 Workload plane: bounded execution, not an unconstrained loop

The workload plane translates an intent into a controlled state transition. Production agents require typed tool contracts, explicit preconditions and postconditions, idempotency keys, transaction boundaries, and compensating actions where rollback is impossible. Asynchronous queues need back-pressure, load shedding, timeouts, retry budgets, and dead-letter handling. Delegated agents inherit an audience- and purpose-bound, attenuated subset of the parent’s authority, time, and spend; they do not widen, replay, or mint new limits. Development, staging, and production identities are separated even if the user interface makes them appear continuous.

This design does not assume a deterministic model. It places deterministic constraints around consequences. A model can select a proposed payee while a policy engine enforces the vendor allowlist and a ledger check proves the postcondition. A model can draft a customer response while destination, disclosure, and send authority remain independently controlled. The model is allowed to be probabilistic where the workflow is prepared to absorb variance.

## 6.3 Control plane: an agent is a delegated principal

Capability, autonomy, and authority are distinct. Authority is implemented through identity, credentials, entitlements, purpose, limits, expiry, separation of duties, and revocation. Current NIST work on software and AI-agent identity describes this as an open standards and adoption problem rather than a solved architecture [National Cybersecurity Center of Excellence, 2026]. A production record should identify the initiating human or service, the accountable sponsor, each on-behalf-of delegation, the exact action authorized, and the policy decision at the system that owns the state. A prompt instruction is not an authorization decision.

Agents compress the distance between content and consequence. The same system can read an email, retrieved page, document, ticket, code comment, or agent message and then act with delegated credentials. We propose representing security exposure through an **authority surface**: privilege breadth, data reach, tool breadth, external reach, delegation depth, action rate, reversibility, and time to selective revocation. These dimensions are a configuration record, not a validated scalar risk score. AI-seat count and benchmark capability are weak proxies for that surface. A lower-capability agent with broad write authority can create more immediate institutional risk than a stronger model confined to a sandbox.

The threat model includes hostile users, malicious insiders, compromised software-as-a-service or tool servers, poisoned documents and memory, malicious model artifacts, provider compromise, cross-tenant leakage, and propagation through peer agents. Email, retrieved pages, code comments, tool descriptions, tool output, memory, and agent messages are untrusted data even when they arrive through an authorized connector. The fundamental prompt-injection problem is that the system processes instructions and adversarial content in the same semantic channel; the practical response is least privilege, trusted tool boundaries, confirmation for consequential ambiguity, and deterministic limits, not a claim that prompting has recreated a security perimeter [UK National Cyber Security Centre, 2025].

Controlled benchmarks establish feasibility, not prevalence. AgentDojo jointly measures task utility and indirect prompt-injection resistance for tool-using agents and shows that useful task behavior can coexist with an exploitable instruction boundary [Debenedetti et al., 2024]. Black Hat demonstrations trace attack paths from untrusted repository, web, retrieval, memory, and tool content into privileged actions [Lynch and Harang, 2025]. OWASP, RSAC, and DEF CON materials provide rapidly evolving taxonomies, exploit demonstrations, and practitioner evidence [OWASP GenAI Security Project, 2025, Byun and Gerchow, 2026, DEF CON, 2025]. These materials establish plausible exploit mechanisms

and practitioner attention, not production prevalence or incidence rates.

The supply chain remains ordinary executable software. A 2025 PyTorch advisory showed that a malicious serialized artifact could enable remote code execution despite a loading mode intended to restrict behavior; affected versions and topology determine actual exposure [PyTorch, 2025]. Model provenance, artifact isolation, patching, signing, software and model bills of materials, runtime hardening, and connector lifecycle management therefore belong inside AI security rather than beside it.

## 6.4 Evidence plane: behavioral service levels and governed traces

Technical uptime is necessary and insufficient. An endpoint can be green while retrieval is stale, policy is bypassed, a tool trajectory is wrong, or the system is producing unacceptable state changes. A behavioral service-level-objective (SLO) set should therefore independently bound valid-acceptance rate, false acceptance and rejection, policy conformance, authoritative-postcondition success, defect escape, escalation, latency, recovery, and cost so that a favorable measure cannot compensate for a prohibited failure. DORA’s 2025 survey of 4,867 professionals is consistent with this systems view: a modeled composite AI-adoption score was positively associated with delivery throughput and self-assessed product performance, and also with delivery instability; platform quality, testing, and feedback loops moderated reported outcomes [Google Cloud DORA, 2025]. The survey’s observational and self-reported design does not prove that AI caused either result.

Post-deployment monitoring is itself an immature boundary. NIST AI 800-4, based on practitioner workshops and a literature review, identifies fragmented logging, difficulty scaling human monitoring, expert shortages, and unsettled methods among the barriers; it organizes concerns but does not validate a monitoring stack [Rao et al., 2026]. This makes observability a designed service rather than a background dashboard. Each material signal needs an owner, consequence-linked threshold, response authority, affected-state query, and tested runbook.

Repeated reliability is a separate requirement. In the published tau-bench study, agents completed fewer than half of the studied tasks and the best reported retail baseline passed all eight repeated trials less than 25 percent of the time [Yao et al., 2025]. A separate software-task study found that the estimated task horizon at 80 percent success was four to six times shorter than the 50-percent horizon [Kwa et al., 2025]. These are benchmark-specific historical results, not permanent limits. They show why a one-run demonstration is an unsafe reliability claim.

An evidence trace should correlate the initiating principal and declared intent with model calls, retrieved source identifiers, memory access, policy decisions, tool mutations, human dispositions, and authoritative-state diffs. OpenTelemetry’s generative-AI semantic conventions provide useful emerging attributes, but their development status and the sensitivity of prompts, outputs, and tool arguments require metadata-first capture, access controls, redaction, retention limits, and deliberate content sampling [OpenTelemetry, 2026]. Observability that leaks protected context can create a separate privacy or security incident rather than a control success.

Evidence becomes operational when it feeds a separately authorized, policy-bounded response path: correlate the trajectory, classify consequence, pause the affected workflow, revoke or quarantine the relevant identity, tool, memory, or bundle, discover state changes, reconcile or compensate them, restore a known-good path, validate behavior, and feed the incident into the next release. The telemetry service should not itself hold unrestricted revocation or state-mutation authority; automated containment must be auditable, and evidence must remain available when the primary control service is compromised. This AI-specific sequence is the paper’s proposed synthesis of monitoring and incident-response practice. Observability without those links can explain a failure after the fact while

leaving the organization unable to stop or repair it. Remediation without an evidence-preserving state-reconciliation step can conceal damage rather than recover from it.

Human review is one evidence mechanism, not a gold standard. Reviewers are variable, biased, fatigued, and susceptible to automation bias. High-volume workflows should compare human-only, machine-only, and combined acceptance errors, including false rejection and review latency. Deterministic checks or an independent verifier may be stronger than a ceremonial approval click. The evidence plane should measure the reviewer as part of the system.

## 6.5 Technical-recovery plane: restore authority and state, not only process health

The incident lifecycle is detect, triage, contain, revoke or quarantine, discover affected state, reconcile or compensate, restore, validate behavior, communicate, and learn. NIST guidance calls for defined escalation, deactivation, disengagement, third-party monitoring, remediation, and tested incident plans [Autio et al., 2024]. Google Cloud’s AI/ML reliability guidance similarly treats evaluation, monitoring, graceful degradation, data and model drift, capacity, and recovery as system concerns rather than model accuracy alone [Google Cloud, 2025a]. For an AI workflow, recovery assets include a known-good behavior-bundle manifest, clean retrieval index and memory, selective identity revocation, affected-state inventory, compensating transactions, an alternate provider or safe-continuity path, and a behavioral suite that must pass before authority is reopened. Technical recovery restores authority, state, and service; retained human capability enables takeover; protected recovery opportunity limits the sustained workload imposed on people. None is a substitute for the others.

Replit’s July 2025 account of an agent deleting a production-connected database is a compact boundary case. Rollback restored the database, but production could be broken during the recovery window. Replit subsequently emphasized dev/prod separation, restrictions on production modification during development, visible rollback, and a planning-only mode [Replit, 2025]. The case does not estimate a failure rate. It shows that model obedience is weaker than environment separation, action permissions, checkpoints, and practiced recovery.

Conventional distributed-systems failures also propagate through AI services. Official Google Cloud and OpenAI incident reports describe configuration or shared-backend failures, control dependencies, retries, and capacity effects across multiple services [Google Cloud, 2025b, OpenAI, 2026b]. A multi-model design is not resilient if all routes share identity, retrieval, network, region, quota control, or telemetry. The control and evidence planes are themselves common-mode dependencies and need integrity, availability, and independent recovery. Provider failover is a behavior release: quality, safety, residency, cost, and postconditions must be revalidated.

## 6.6 Runtime economics belong in the architecture

Cost controls should act at request, session, delegated subtree, tenant, project, and period levels. Relevant limits include tokens, elapsed time, tool calls, iterations, delegation depth, concurrency, retrieval fan-out, transactions, and retries. Equation 4 makes routing and model choice subordinate to valid outcomes. A cheaper route that increases correction, incidents, or false acceptance is not cheaper; an expensive model that sharply reduces acceptance labor can be. Research on agent evaluation likewise shows that rankings can change when cost and reproducibility accompany accuracy [Kapoor et al., 2025]. The operational test is a cost–quality–consequence frontier on the organization’s own tasks, including egress and switching costs, not a public price-per-token table.

Average-load planning is especially weak for AI runtime. A 2026 NSDI study characterized 3.54 billion requests across 12 models in one worldwide cloud service. Arrival burstiness and input and output lengths shifted over time; in one workload, 29 of 2,412 clients generated 90 percent of requests

[Xiang et al., 2026]. Those distributions should not be generalized to every organization, but they show why workload class, upper-tail demand, client concentration, quota behavior, and retry amplification belong beside average tokens. The financial-control boundary is also moving quickly: among the 693 respondents answering the AI-spend item in the 2026 State of FinOps survey, 98 percent reported managing AI spend; AI cost management was also the leading desired skill area among 695 respondents [FinOps Foundation, 2026]. The report had 1,192 total respondents. This is community survey evidence, not a population estimate, but it supports treating allocation and value attribution as operating functions rather than annual procurement exercises.

## 6.7 Conditional forecast: scale acceptance capacity before it binds

Equation 6 operationalizes the theory as a proposed pre-emptive forecast. A potential planning window is suggested by the adoption gap: the 2026 AI Index reports broad organizational use while agent deployment remains in the single digits across most surveyed business functions [Stanford Institute for Human-Centered Artificial Intelligence, 2026]. Anthropic’s vendor telemetry reports millions of interactions in which the extreme tail of Claude Code turn duration increased while experienced users used automatic approval more often; actual runtime is not the human-equivalent task horizon measured by capability benchmarks [Anthropic, 2026]. An OpenAI-authored preprint using production Codex telemetry reports more than fivefold active-user growth in the first half of 2026 and that more than 10 percent of users managed at least three concurrent agents during some week [Johnston et al., 2026]. Both telemetry sources are product-, population-, and period-specific; they measure use patterns rather than accepted value or a universal capability clock. They identify signals that an organization can measure locally before authority and volume expand.

Table 9 presents conditional states rather than an AGI timetable. Activation requires a joint pattern of local indicators; no single public benchmark, seat count, runtime, or calendar date is sufficient.

**Table 9:** Proposed conditional operating states and capacity implications

| Proposed operating state; evidence status             | Local signals and release gates                                                                                                                                                                                 | Hypothesized bottlenecks and failure modes                                                                                                                            | Scale before crossing                                                                                                                                                                                                                                                                                         |
|-------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Bounded delegation; observed, population-specific     | Mostly single trajectories; short actual runs; material state change remains separately authorized; rising use but limited delegation                                                                           | Domain acceptance; data quality; evaluation coverage; support; cost visibility; human adoption and recovery load                                                      | AI platform and FinOps; data and knowledge owners; evaluation and domain-review capacity; security architecture; role-specific enablement and protected recovery                                                                                                                                              |
| Supervised parallel production; emerging              | Sustained growth in candidates per root, trajectories per candidate, cumulative agent-hours, delegated tool breadth, privilege-weighted write actions, and upper-tail concurrency                               | Agent identity and policy; inference bursts; trajectory evidence; specialist review; incident containment; state reconciliation; human attention and on-call capacity | Agent-platform engineering; IAM and security; SRE and observability; evaluation operations; red teams; workflow owners; staffed rotations and workload controls                                                                                                                                               |
| Persistent semi-autonomy; conditional stress scenario | Local repeated reliability meets consequence thresholds; uninterrupted runtime spans the operating interval; state-changing delegation compounds; trace, revocation, reconciliation, and continuity drills pass | Correlated failure; stale context; machine-speed incidents; exception adjudication; provider concentration; institutional memory and continuity                       | Machine-enforceable authority envelopes; independent automated verification; coverage matched to operating hours, consequence, and recovery objectives, including 24/7 where agents act continuously; model-risk assurance; provider and compute resilience; apprenticeship and manual or degraded continuity |

The dashboard should preserve the causal chain: root requests, candidates per root request, trajectories per candidate, visits and service time per visit by boundary resource, queue and tail latency, privilege-weighted state changes, trace completeness, false acceptance, defect escape, containment and reconciliation time, cost per valid outcome, and human recovery indicators. Candidate long-lead assets include authoritative data, nonhuman identity and policy engineering, independent evaluation, incident muscle, domain judgment, and the human capability to challenge and continue. A managed provider can supply parts of this spine, but it cannot eliminate the need to know who owns the inherited control, what evidence it emits, and how the organization continues when it fails.

## 7 Venture capital, startups, and investment concentration

Venture capital appears twice in the AI race: as an allocator of capital toward models and infrastructure, and as an organization using data and AI to allocate capital. The two roles expose the same boundary problem. More compute can expand the frontier; more candidates and more polished information can also saturate selection and concentrate attention around what is easiest to score.

### 7.1 Data-driven investing scales search more readily than novelty

The strongest direct evidence is [Bonelli \[2026\]](#). The study combines global Crunchbase, employee, patent, exit, and Preqin data and identifies a VC as data-driven when it hires its first employee whose role indicates pre-investment data infrastructure or algorithms. After that hire, adopting firms increased annual deal count and assets under management by roughly 30 percent more than other VCs. Selection of startups resembling previously funded businesses rose sharply; the probability that a selected startup both survived and received follow-on financing increased by 2.5 percentage

points, about 5 percent relative to a 52 percent unconditional rate. Yet the probability of the paper’s rare major-success outcome—IPO or profitable acquisition—fell by 2.5 percentage points, about 17 percent relative to its baseline, and portfolio companies later produced fewer patents and citations. Fund performance did not improve significantly.

The observational adoption result is not itself causal. The paper’s stronger causal argument uses a quasi-experimental, continuous-treatment difference-in-differences design based on VCs’ pre-shock exposure to the AWS-induced startup boom; it still depends on parallel-trends and treatment-design assumptions. Textual similarity is a proxy for novelty, and the survival-and-follow-on composite is not investor return. The bounded conclusion is important: data infrastructure can scale screening and improve common-outcome selection while shifting the portfolio toward backward-legible opportunities. In a power-law domain, optimization for the frequent label can reduce exposure to the rare objective.

SignalFire’s Beacon illustrates the operating form. The firm reported a system covering about 80 million companies, 600 million people, and 40 datasets, with language models used for similarity, descriptions, tags, and title normalization [Kirmos, 2023]. It also explicitly retained human outreach, reading, and final investment judgment. This is credible evidence of a production research layer, not of superior fund returns or an identified optimum. The observed deployment boundary is upstream; the paper’s testable sweet-spot hypothesis is that machines can expand and normalize the searchable universe while partners retain the consequential thesis, relationship, valuation, and commitment.

## 7.2 Generative polish changes the screening equilibrium

When generation becomes cheap, a well-formed artifact can carry less information about the underlying producer. In experiments spanning startup investing and hiring, sender access to ChatGPT improved presentation but reduced evaluator screening accuracy by 4–9 percent on average and increased demand for costlier corroborating signals by 6–9 percent [Cowgill et al., 2026]. Some non-English-speaking senders became easier to assess, and not every error specification was significant. The result is not that AI-assisted founders are worse. It is that style becomes less diagnostic when it is widely subsidized.

For venture firms, this converts first-pass efficiency into a new acceptance tax: primary customer evidence, product telemetry, reference calls, technical reconstruction, data-room provenance, and observed founder reasoning become more valuable. For fund managers and portfolio companies, unsupported AI claims also create disclosure risk. The SEC’s 2024 enforcement actions against two investment advisers establish that materially false or misleading claims about AI use can be ordinary securities-law violations; they do not estimate market-wide “AI washing” [U.S. Securities and Exchange Commission, 2024]. The acceptance boundary for an AI investment thesis therefore includes evidence of the deployed workflow and its operating economics, not merely model access or an AI label.

## 7.3 The capital race is concentrated at the frontier

The OECD estimates that AI companies received \$258.7 billion in global venture investment in 2025, 61 percent of all global VC; 73 percent of AI investment value occurred in rounds above \$100 million, and infrastructure and hosting received \$109.3 billion, or 42 percent [OECD, 2026]. PitchBook and NVCA report that U.S. AI deal value reached \$355.9 billion in the first half of 2026, 86 percent of total value, while AI represented 43.2 percent of deal count. Across all U.S. VC, megadeals accounted for 87.5 percent of first-half deal value [PitchBook and National Venture Capital Association, 2026]. Definitions, geography, periods, and late deal revisions differ, so the figures should not be merged. Both show that the financing race is much more concentrated than a count of AI products suggests.

Supplier filings expose concentration within the infrastructure layer. CoreWeave’s 2025 prospectus reported Microsoft at 62 percent of 2024 revenue and its two largest customers at about 77 percent [CoreWeave, Inc., 2025]. Cerebras reported that G42 supplied 87 percent of first-half 2024 revenue and 97 percent of hardware revenue [Cerebras Systems Inc., 2024]. These are company- and period-specific disclosures, not proof that AI infrastructure is generally unstable. They show that rapid scale can coincide with dependence on one counterparty; Cerebras additionally disclosed prepayments and purchase commitments that reduced working-capital needs.

At the model layer, API-market evidence describes many providers, rapid product turnover, large price declines, and different leaders by use case, even as most buyers concentrate use on a small number of models [Demirer et al., 2025]. This combination makes portability economically meaningful but not free. A second provider does not remove correlated cloud, identity, network, data, or accelerator dependencies. Conversely, a managed provider may deliver security and reliability that a small firm cannot reproduce. The durable organizational asset need not be an internally built operating substrate. It may be the firm’s workflow definitions, permission mappings, evaluation cases, outcome history, incident-derived regressions, and recovery knowledge—artifacts that improve both vendor governance and migration.

## 7.4 An operating-diligence screen

For an investor evaluating an AI startup or an enterprise evaluating a vendor, model benchmarks should be the beginning of diligence. The acceptance-boundary view asks:

1. What authoritative state changes, and who remains accountable for them?
2. Which task distribution and consequence tier were actually evaluated, with what repeated reliability?
3. What are valid-acceptance, false-acceptance, escalation, defect-escape, and realized-outcome rates?
4. Which data, identity, tool, cloud, model, customer, and financing dependencies are common-mode?
5. What human acceptance and exception labor is required per valid outcome after the adoption cohort matures, and has faster production increased work intensity or reduced recovery?
6. Which shared resource has the shortest acceptance-capacity runway, and what intervention is funded before it binds?
7. Can authority be selectively revoked and affected state discovered, reconciled, and safely continued?
8. Which behavior-bundle evidence and workflow data remain portable if price, quota, model, or provider changes?
9. Does the optimization target preserve exploratory option value and rare-outcome discovery, or only improve common labels?

This screen does not favor infrastructure over applications, build over buy, or humans over machines. It favors an explicit mapping from claimed capability to accepted value, bounded authority, unit economics, and recoverability.

## 8 Nine propositions and falsifiers

The framework is useful only if it produces observations that could be wrong. The propositions below specify a unit, a measurement design, and a direct falsifier. They are not estimated coefficients.

**P1: Acceptance-capacity runway predicts resource-SLO breach.** When AI raises candidate or action demand faster than the service capacity of the bottleneck acceptance resource, shorter forecast runway in Equation 6 should precede earlier breach of the resource’s declared SLO.

For congestible resources approaching saturation, it should also precede nonlinear growth in work in process and tail lead time even if generation time falls. *Test*: preserve rolling-origin forecasts; instrument root requests, candidates per root, trajectories per candidate, boundary visits, service per visit, committed capacity, utilization, rework, abandonment, and defect escape; and compare runway with baselines using seats, tokens, or lagged load. Record interventions so a prevented breach is not counted as a failed forecast. *Falsifier*: runway is persistently miscalibrated or does not rank first SLO breach better than the simpler baselines. The load mechanism is separately weakened if materially higher demand produces proportional validly accepted throughput without worse latency, rework, abandonment, or defect escape.

- P2: Stage-specific authority configuration.** Authority should differ across acquisition, analysis, selection, and execution, with prespecified positive interactions between autonomy and verifiability or reversibility, and negative interactions between autonomy and consequence or weak recovery. *Test*: randomize or phase stage-specific authority within preregistered task and consequence strata while holding model and task mix fixed. *Falsifier*: the interaction estimates are practically null and one fixed profile yields equal or higher valid accepted value at every stage under the prespecified strata. Monotonic improvement for sandboxed candidate work does not falsify the proposition for state-changing authority.
- P3: Behavior-bundle effects beyond model scores.** After the ex ante, workflow-local capability threshold defined in Section 3.1 is crossed, retrieval, data freshness, tool contracts, permissions, validators, routing, and recovery configuration will explain material held-out production variance beyond general benchmark scores. *Test*: run factorial ablations and versioned canaries on local valid-acceptance and defect outcomes. *Falsifier*: model identity or benchmark score predicts those outcomes as well as the full configuration across relevant changes, while manifesting adds no reduction in change failures or investigation time.
- P4: Shared operating-substrate complements moderate organizational value.** Beyond bounded tasks, access should produce larger durable gains where reusable identity, governed data, policy, evaluation, evidence, cost, and recovery primitives are supplied as a common service while workflow semantics, exceptions, and residual-risk decisions remain with accountable domains. This extends the complementarity and productivity-J-curve literature and is consistent with recent organizational evidence [Bresnahan et al., 2002, Brynjolfsson et al., 2021, Chatterji et al., 2026, Google Cloud DORA, 2025]. A qualitative large-enterprise case similarly frames governance as an architectural learning mechanism, not a final approval layer [Dux et al., 2026]. *Test*: preregister the inherited-control and domain-ownership configuration, then use multi-site or stepped-wedge rollouts to estimate its interaction with access on valid outcomes, platform queue time, duplicated controls, incidents, and recovery at six and twelve months. *Falsifier*: the configuration adds no explanatory value, or its delay and common-mode failures offset its benefits while access alone produces noninferior durable end-to-end value outside bounded tasks.
- P5: Authority amplifies system risk independently of capability.** For a given task-success rate, expected and tail loss will increase with credential scope, action irreversibility, delegation depth, external reach, and weak transaction limits. *Test*: red-team matched configurations with identical models and progressively broader authority, measuring unauthorized state change and containment. *Falsifier*: authority scope adds no predictive information after model capability and task difficulty are controlled, or a model-level safeguard blocks all relevant action paths without external enforcement.
- P6: Recovery evidence predicts resilience better than nominal failover.** Systems that practice selective revocation, affected-state discovery, reconciliation, known-good restoration, and

degraded continuity will recover consequential service more reliably than systems that test end-point or provider failover alone. *Test*: inject matched model, tool, identity, data, and control-plane faults and compare time to safe, reconciled service. *Falsifier*: untested provider substitution yields equivalent behavioral conformance, state integrity, and recovery time across relevant faults.

**P7: Workload controls and active practice address different human risks.** After assisted performance is matched, organizational workload controls should reduce sustained work intensity and fatigue-sensitive error while increasing protected recovery opportunity and perceived control; active-practice mechanisms should preserve unaided error detection, task completion, calibration, and takeover. Neither intervention should substitute for the other. *Test*: with worker participation and privacy safeguards, randomize or phase workload control and active practice as separate factors, preregister clinically non-diagnostic cohort outcomes, minimum important differences, and noninferiority margins, and test interaction and spillover effects. *Falsifier*: workload controls do not improve the prespecified recovery-opportunity outcomes, active practice does not preserve the prespecified unaided outcomes, or either factor reliably produces both effects without hidden transfer of work to another group.

**P8: Multi-agent returns are conditional.** Coordinated multi-agent systems should outperform a compute-matched single agent or independent best-of- $N$  baseline only when task decomposition, useful error diversity, and verification quality exceed coordination and tool overhead. Before allocation, decomposability should be scored from the task-dependency graph, error diversity from baseline error correlations, and verification quality from independently labeled acceptance error. A NeurIPS trace corpus identifies recurring system-design, inter-agent alignment, and verification failures [Cemri et al., 2025]; a broad preprint reports gains on decomposable tasks and large losses on sequential planning under standardized configurations [Kim et al., 2025]. Simulated AI organizations also show that composition can change utility and alignment, but results are model- and scenario-dependent [Shen et al., 2026]. *Test*: compare compute- and token-matched architectures within preregistered strata while measuring communications, tool calls, latency, cost, and valid acceptance. *Falsifier*: the prespecified interactions are practically null and multi-agent designs consistently dominate strong matched baselines across sequential, tool-heavy, and nondecomposable tasks.

**P9: Data-driven screening can trade off against novelty.** In high-volume, power-law selection domains, data-driven screening without novelty-preserving objectives or sampling should improve common, observable intermediate outcomes while reducing exploration [Bonelli, 2026]. *Test*: in venture or comparable innovation funnels, randomize standard data-driven screening against a preregistered novelty-preserving arm, retain rejected candidates, and track sourcing coverage, similarity, false negatives, follow-on, rare major outcomes, patent novelty, partner overrides, and returns. *Falsifier*: in the prespecified no-preservation arm, scaled screening improves both common selection quality and rare-outcome discovery without narrowing the candidate distribution. The AI-pitch evidence in Cowgill et al. [2026] is an adjacent signal-degradation mechanism, not an estimate of this treatment.

## 9 Implications for operators, boards, and investors

## 9.1 Operators: scale the shared substrate before the boundary binds

The first implementation unit is a consequential workflow, not a model contract. The owner should define the authoritative postcondition, consequence tier, acceptance and validity tests, delayed outcome, authority profile, behavior-bundle manifest, safe-continuity path, and unit-cost denominator. Baseline human-only false acceptance, false rejection, elapsed time, rework, and defect escape before changing the workflow. This prevents pre-existing bureaucracy from being misclassified as an AI tax and makes obsolete gates visible.

The operational dashboard should start with root requests, candidates per root, trajectories per candidate, boundary visits and service per visit, observed acceptance, valid acceptance, defect escape, accepted-throughput lead time, human minutes, exception distribution, fully loaded cost, authority-surface change, trace completeness, tested recovery time, and acceptance-capacity runway. Human-recovery measures should include sustained concurrency, alerts and interruptions, after-hours load, protected recovery, perceived control, and fatigue-sensitive error at an aggregate cohort level. Prompt counts, token counts, licenses, or generated artifacts can diagnose use; they are not primary value metrics. Capacity should be managed at the resource-class bottleneck, and incidents or material overrides should become versioned regression tests.

Central platforms are useful when they make identity, governed data access, policy, evidence, evaluation, cost control, rate shaping, and recovery reusable. They become harmful when the common layer is opaque, slow, surveillant, or a single failure domain. A product team should be able to prove which controls are inherited, which remain local, and how both layers recover. The central function should publish service objectives and lead times so domain teams can forecast runway rather than discover scarcity in a launch queue. Build-versus-buy follows the same logic: choose the arrangement that produces stronger evidence, state integrity, portability, response, and sustainable operating load at the required consequence, not the one with the largest internal component count.

## 9.2 Boards and executives: govern authority, not enthusiasm

Board oversight should distinguish portfolio exposure from operating readiness. A useful portfolio view shows, by consequence tier, the number of deployed workflows; maximum authority and external reach; valid-acceptance and defect-escape trends; the minimum acceptance-capacity runway; concentration by provider, cloud, data source, and specialist reviewer; material incidents and near misses; technical and human recovery indicators; fully loaded economics; and retained-capability tests. Aggregate adoption can accompany this view but cannot replace it.

Decision rights should be explicit. One executive sponsors cross-plane integrity, and each shared service has a named owner; business owners define value and postconditions; security and identity owners define permissible authority; data owners define provenance and rights; reliability owners define service and recovery evidence; finance defines fully loaded economics; People and occupational-safety owners govern sustainable job design; and accountable workflow executives accept residual exposure. Independent assurance challenges consequential configurations. This is not a demand for committee approval on every low-consequence action. Deterministic policy and automated evidence should remove ritual review where they outperform it.

## 9.3 Investors: underwrite the operating boundary

The investable question is not whether a company uses a frontier model. It is whether its valid accepted outcomes improve as volume, customers, tools, and authority expand. Diligence should reconcile claimed gross margin with inference, data, evaluation, human acceptance, support, incident,

recovery, and provider-switching costs. It should inspect customer and infrastructure concentration, evidence portability, false-negative and false-acceptance denominators, the recovery path for corrupted state, reviewer concentration, after-hours operating load, and the production of future experts. In vertical applications, proprietary advantage may live in outcome-linked workflow data and exception handling rather than prompts. In infrastructure, growth should be evaluated alongside capex, debt, counterparties, energy, supply chain, and common-mode dependencies.

VC firms should apply the same discipline internally. Sourcing systems can widen the funnel and reduce clerical search; investment committees should audit whether ranking compresses novelty, whether AI-assisted pitches have changed signal value, whether primary diligence and partner attention have scaled with candidate production, and whether speed has merely extended the workday. Portfolio operating teams can provide a reusable control spine across small companies—identity patterns, security baselines, evaluation, observability, FinOps, recovery drills, and vendor-concentration analysis—without recreating a large-enterprise department inside every startup. A model that helps find the median good company can still be poorly aligned with a fund whose return depends on rare outliers.

## 9.4 AGI scenarios as an operational stress test

Conceptual work proposes scales for increasingly autonomous or superintelligent systems [Genewein et al., 2026, Chojecki, 2025], while scenario essays propose compressed development timelines [Aschenbrenner, 2024, Kokotajlo et al., 2025]. None supplies an empirical timetable for this paper. Their operational value is as a stress test: what if capability, action speed, delegation depth, and candidate production rise faster than procurement, identity governance, review staffing, incident learning, and infrastructure expansion?

The runway construct turns those narratives into capacity shocks rather than dates: vary root requests, candidates per root, trajectories per candidate, visits and service per visit, authority breadth, provider availability, and expansion lead time, then observe which resource loses safe headroom first. The stress result is conditional on the scenario and local task distribution; it is not a probability of AGI.

Under that stress, the correct response is not to freeze today’s human approval pattern or make people absorb machine-speed demand. It is to convert policy into machine-enforceable limits, make authority revocable, instrument trajectories and state, expand automated independent verification, preserve technical and human continuity, shape arrival rates, and remove handoffs that do not produce proof. If future systems become reliably self-verifying and self-recovering across the local task distribution, the boundary should move. The framework is designed to register that improvement through lower  $d_j$ , higher  $p$ , lower defect escape, longer runway, and faster tested recovery rather than defend a permanent role for manual review.

## 9.5 Counterarguments and boundary conditions

Several alternatives could weaken the thesis. First, apparent boundary load may be legacy bureaucracy. If a gate neither reduces false acceptance nor supplies accountability, recovery, or learning, AI-enabled redesign should remove it. Second, people are imperfect validators; independent deterministic or machine checks may dominate human review. Third, a stateful autonomous system can reduce handoff and context costs rather than increase them. Fourth, strict accepted-outcome optimization can suppress exploration where truth is delayed and rare options matter. Fifth, a managed vendor may operate the control and evidence planes more safely than an internal team. Sixth, a central platform can become a bureaucratic or technical bottleneck whose common-mode risk exceeds the

fragmentation it replaced. Seventh, ecological, energy, psychosocial, and social constraints may bind upstream of the measured acceptance resource [Bhardwaj et al., 2025].

These alternatives do not disappear inside the framework. They change the measured terms and feasible region. The central claim would be substantially weakened if model capability and access alone predicted sustained organizational value across heterogeneous environments; if configuration, authority, validation capacity, and recovery added no explanatory power; or if organizations could repeatedly scale consequential AI without explicit state, evidence, and continuity controls while maintaining quality and bounded loss.

## 10 Limitations and research agenda

This paper is a structured integrative review, not a systematic review, meta-analysis, or causal test of the boundary model. The search was purposive and iterated as constructs developed. No comprehensive count of all retrieved records was maintained, no second human coder independently screened sources, and no inter-rater reliability statistic is available. Appendix A reports the search families, inclusion logic, and coding fields so the scope is inspectable without implying PRISMA-level reproducibility.

The empirical base is selected in several ways. Successful deployments are easier to publish than failed pilots; incidents with responsible disclosure are easier to observe than quiet near misses; digital workflows generate more telemetry than tacit or physical work; and named companies may have unusually strong platforms, incentives, or vendor support. Vendor cases establish reported use, not causal value. Filings are authoritative for disclosure but still rely on company definitions. Experiments often use short tasks, selected workers, and dated model snapshots. Working papers and 2026 preprints may change. The workforce evidence supports monitoring work intensity, autonomy, and recovery opportunity; it does not establish that enterprise AI causes a clinical mental-health condition or burnout. Sensitive workforce measurement can itself create surveillance and selection bias.

The proposed constructs also embed judgment. Quality, sufficient evidence, consequence, option value, and maximum credible harm cannot always be reduced to objective labels. An organization can define a weak validity test and appear efficient. Delayed truth in investing, strategy, science, and innovation makes false rejection especially hard to see. Authority and legal legitimacy vary across jurisdictions. The queueing equations are fluid heuristics; resource demand, capacity, task mix, and acceptance probability are endogenous and nonstationary. The autonomy optimization is a decision aid, not a social-welfare function.

The boundary may move quickly. Better models, tool learning, uncertainty estimation, mechanistic evaluation, and automated verification could lower acceptance demand. Conversely, larger action spaces, agent-to-agent communication, hidden provider updates, more sophisticated attacks, and accumulated dependency could raise it. General benchmarks are necessary for comparative science but cannot substitute for versioned local evaluation. Frontier safety frameworks address model-level capability and deployment risk [Google DeepMind, 2026]; the present framework addresses a different scale and should be tested for complementarity rather than treated as a replacement.

The highest-value research program would build a multi-organization dataset that follows candidates through  $D$ ,  $Z$ , and delayed  $Y(h)$ , with versioned behavior bundles, authority profiles, review-resource classes, forecast runway, human touch and recovery measures, cost, and incidents. Nine study designs are particularly useful:

1. stepped-wedge changes in authority profile within the same workflow;
2. out-of-sample runway forecasts and load experiments that estimate congestion effects on human

- and automated validation;
- 3. factorial ablations of retrieval, tool contracts, permissions, evaluators, and routing;
- 4. worker-participatory longitudinal studies of workload controls, protected recovery, apprenticeship, and assisted versus unaided performance using privacy-preserving measurement;
- 5. compute-matched single-agent, sampling, centralized, decentralized, and hybrid agent comparisons;
- 6. paired human-only, machine-only, and combined acceptance-error studies;
- 7. fault injection across model, data, tool, identity, provider, and control-plane dependencies;
- 8. managed-versus-internal platform comparisons using valid outcomes, incidents, cost, and portability; and
- 9. venture and innovation studies that retain rejected candidates and measure rare outcomes, novelty, and option value over long horizons.

A common measurement schema would allow the claims in Section 8 to be estimated rather than illustrated. Negative and null results are especially important: they reveal where the boundary framing adds no predictive value or where simpler task-fit and organizational-complement theories are sufficient.

## 11 Conclusion

The AI race is usually narrated as a race between models. Organizations experience a second race between candidate production and the capacity to create valid, authorized, observable, recoverable state change. The acceptance boundary makes that downstream race measurable. It offers a unifying interpretation of how causal task-level gains can coexist with slower expert work, review saturation, weakened screening signals, unstable delivery, and uncertain firm-level return: the observations occur at different points in a production system.

The evidence supports a conditional boundary thesis. Capability remains decisive until a configured system can perform the task. Beyond that threshold, valid acceptance can become the weak link. The feasible autonomy region is widest where objectives and postconditions are measurable, feedback is fast, verification is independent and cheap, actions are bounded or compensable, tools and permissions are narrow, shared resources have headroom, and technical and human continuity are practiced. It narrows where truth is delayed, context is tacit, loss is irreversible, credentials are broad, errors are correlated, or scarce experts must absorb both production and proof.

This is not a case for permanent human vetoes or a single approval bureaucracy. Humans are fallible, review can become theater, and legacy gates can destroy value. For consequential workflows, the proposed pattern is a logically shared, failure-isolated operating substrate that makes identity, governed data access, policy, evaluation, evidence, cost control, and technical recovery reusable while accountable domain teams retain workflow meaning, postconditions, exceptions, and residual-risk decisions. Security exposure should be assessed through the authority surface; seat count alone is insufficient because content becomes materially more dangerous when an agent can transform it into privileged action. Observability becomes useful when it supports a separately authorized path for selective revocation, affected-state discovery, reconciliation, safe restoration, and learning.

The acceptance-capacity runway turns that architecture into a forward operating test. It asks which shared resource will breach a locally defined service or safety objective before added capacity can be designed, procured, staffed, trained, validated, and released. Capacity can be expanded through automation, stronger proofs, demand shaping, process redesign, managed services, or people. It cannot be manufactured by treating worker attention as elastic. Sustainable scaling therefore includes protected recovery opportunity, manageable exception and incident load, worker participation, and

preserved unaided capability; a recovery prompt is at most a secondary aid, not a substitute for work design.

Nor is this an AGI timetable. It is a conditional forecast of operational pressure. If generation, parallelism, and delegation accelerate by orders of magnitude, identity, data rights, verification, observability, state reconciliation, finance, and retained judgment must become more machine-readable, selective, and scalable. If future systems reliably reduce service demand, false acceptance, and recovery time, the measured boundary should move; authority may expand only after reauthorization within the workflow’s legal, consequence, reversibility, and residual-risk constraints. The framework is weakened if seats, tokens, or model scores predict durable value and bounded loss as well as the proposed operating measures.

Venture capital makes the asymmetry visible in both directions. Capital is concentrated upstream in large rounds and infrastructure, while operating value depends on slower complementary assets. Inside the investment process, analytics can widen sourcing and improve common-outcome selection while compressing novelty and making polished narratives less diagnostic. The durable edge is therefore not AI access alone. It is the ability to connect changing models to proprietary workflow evidence, bounded decision rights, primary diligence, recoverable operations, and a learning system that remembers failures.

The practical claim is demanding but testable: scale valid accepted outcomes, not artifacts; forecast acceptance-capacity runway before new demand and authority exhaust it; release and evaluate the behavior bundle, not the model name; make shared identity, policy, evidence, and technical recovery reusable while retaining domain ownership of outcomes and residual risk; grant authority as an expiring profile, not a side effect of tool access; observe trajectories and state, not uptime alone; recover by revoking, discovering, reconciling, and validating, not merely restarting; and preserve both the unaided human capability and sustainable workload needed for challenge and continuity. Capability begins at the frontier. Institutional readiness is decided at the acceptance boundary.

## A Evidence coding protocol

### A.1 Scope and search process

The evidence cutoff was August 21, 2026. The review proceeded in four passes: (1) concepts and cases raised by the author’s prior operational writing and supplied materials; (2) targeted scholarly and official-source searches; (3) backward and forward source chaining around retained papers; and (4) an adversarial pass for counterexamples, alternative explanations, primary-source replacement, and metadata verification. The principal locations were arXiv and conference proceedings; NBER and publisher pages; institutional research repositories including Stanford, Google/DeepMind, Microsoft, and major universities; SEC EDGAR; OECD, U.S., U.K., and E.U. public data; NIST and maintainer advisories; and official DEF CON, Black Hat, and RSAC records.

Search strings and retrieved-result counts were not logged comprehensively, and discovery records that failed early relevance checks were not retained. The work therefore cannot support a flow diagram or exhaustive-recall claim. The bibliography identifies retained sources, not the full discovery universe.

### A.2 Eligibility and source replacement

A source was retained when it supplied at least one of: a causal or longitudinal design with a defined outcome; a peer-reviewed benchmark with enough information to locate the measured behavior; an official filing, government statistic, standard, advisory, or postmortem; a named deployment with both

**Table 10:** Search families used in the structured synthesis

| Domain                        | Illustrative query families                                                                                                                                                | Target evidence                                                                      |
|-------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| Capability and agents         | scaling laws; test-time compute; task horizon; tool agents; repeated reliability; multi-agent failure and scaling                                                          | Model/task limits, architecture interactions, evaluation definitions                 |
| Work and organizations        | generative AI field experiment; software productivity and release; organizational AI adoption; team design; work intensity; psychosocial risk; deskilling; automation bias | Causal task effects, attenuation, complements, human recovery, retained capability   |
| Infrastructure and operations | AI agent identity; MCP/tool security; GenAI observability; SLO; RAG failure; AI incident; rollback; workload forecast; FinOps                                              | Control design, exploit and incident mechanisms, capacity, recovery, cost            |
| Venture and startups          | data-driven VC; AI-native firm; algorithmic sourcing; AI pitch screening; AI VC investment; infrastructure concentration                                                   | Search and selection effects, org design, financing and dependency                   |
| Named deployments             | company name plus AI assistant, annual report, filing, evaluation, customer support, course production, code review                                                        | Location of automated stage, acceptance event, retained authority, disclosed outcome |

task and decision boundary; or a conceptual model necessary to position the contribution. Vendor or company material was eligible for feasibility and self-reported practice, not causal return. Scenario essays were eligible only as stress-test context.

Excluded from core claims were anonymous claims without method, unsourced productivity figures, secondary commentary when the primary filing or paper was available, market totals without geography/period/denominator, vulnerability claims without affected scope, and conference material treated as prevalence. When a peer-reviewed version replaced a working paper, the published version was preferred. User-provided documents supplied hypotheses and terminology but were not granted evidentiary weight by origin.

### A.3 Coding fields and inference rule

Each major source was coded for: identifier and version date; publication status; funder or interested party; design; setting and sample; unit of analysis; treatment or system configuration; measured outcome and denominator; uncertainty; acceptance stage; authority; consequence; principal result; alternative explanations; external-validity limit; supported manuscript claim; prohibited inference; and potential falsifier. Company cases additionally recorded whether a number came from a filing, regulator, government report, company release, or vendor page.

The governing inference rule was: state what the source directly measured, preserve its denominator and unit, label status and interested-party provenance, stop at the design boundary, and identify at least one observation that could weaken the interpretation. Source types in Table 1 summarize status and claim fit; they are not a universal quality ranking. The manuscript was prepared with AI assistance for source discovery, synthesis support, drafting, bibliographic-metadata verification, and separate adversarial review passes focused on academic novelty and operations architecture. No second human coder independently screened the corpus; no inter-rater reliability statistic is available.

The named author is responsible for the final claims and citations.

## B Accepted-outcome scorecard

The scorecard is a measurement template, not a compliance standard. It should be completed for one workflow and behavior-bundle release, then linked to execution evidence and reviewed at a cadence appropriate to the consequence tier.

**Table 11:** Accepted-outcome scorecard

| Domain                       | Required definition                                                                                                                                     | Primary measures or evidence                                                                                                               | Release question                                                                                          |
|------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|
| Purpose and owner            | Named workflow, user, business objective, accountable sponsor, system of record, expiry/review date                                                     | Inventory record; business baseline; owner attestation                                                                                     | Is there a durable problem and an owner of the outcome?                                                   |
| Outcome layers               | Observable decision $D$ ; contemporaneous validity $Z$ ; delayed value or defect $Y(h)$                                                                 | Acceptance log; validation result; later outcome join; false acceptance and rejection                                                      | Can acceptance, validity, and later value be distinguished?                                               |
| Task distribution            | Normal, tail, adversarial, multilingual, inaccessible, and abstention cases; version date                                                               | Held-out local evaluation; pass-at- $k$ or repeated-volume reliability; confidence interval                                                | Does evidence match the tasks and repetition the service will face?                                       |
| Consequence                  | Effects on money, rights, confidentiality, safety, external commitments, reputation, and authoritative records                                          | Tier/vector; maximum credible harm; prohibited action list; tail threshold                                                                 | Are non-compensable constraints explicit?                                                                 |
| Authority profile            | Goal, planning, data, tool, action, external communication, persistence, delegation, spend, and transaction limits                                      | Identity and delegation graph; policy tests; separation of duties; expiry; revocation result                                               | Is authority narrower than capability and independently enforceable?                                      |
| Behavior bundle              | Model/route, application/runtime, retrieval/memory, data/index, tools, identity, policy/budgets, evaluation/telemetry, canary, fallback                 | Signed or controlled manifest; resolved execution record; dependency and advisory inventory                                                | Can a release and affected executions be identified?                                                      |
| Data rights and quality      | Source authority, purpose, lineage, freshness, tenant ACL, residency, retention/deletion, poisoning response                                            | Source and index identifiers; access tests; stale/duplicate rate; deletion and quarantine test                                             | Does retrieval preserve rights and authority rather than merely add context?                              |
| Boundary capacity and runway | Resource-class arrival, service demand, capacity, utilization, variability, priority, rework, abandonment, SLO threshold, and expansion lead time       | Queue telemetry; work in process; median and tail lead time; predictive load; shortest runway; intervention record                         | Will required validation remain performed through the time needed to add safe capacity?                   |
| Validity and quality         | Quality, authority, policy, evidence, and postcondition predicates; delayed defect definition                                                           | Valid-acceptance rate; false acceptance/rejection; defect escape; override and escalation                                                  | Is the metric harder to game than generated volume?                                                       |
| Security and privacy         | Assets, adversaries, trust boundaries, entry points, loss modes, telemetry sensitivity                                                                  | Threat model; red-team results; negative authorization tests; data leakage and injection tests                                             | Can untrusted content cause an unauthorized consequence?                                                  |
| Economics                    | Model, data, infra, evaluation, security, human, support, incident, recovery, egress, and switching cost                                                | Fully loaded cost per valid outcome; value- and quality-adjusted frontier; sensitivity analysis                                            | Does value survive review, exceptions, and dependency stress?                                             |
| Technical recovery           | Stop, revoke, quarantine, affected-state discovery, reconciliation, restoration, behavioral validation, degraded/manual continuity                      | Timed fault-injection drill; RTO/RPO; state diff; known-good manifest; reopen record                                                       | Can the organization restore safe state and service, not only restart a process?                          |
| Human recovery               | Sustainable workload, protected recovery opportunity, decision latitude, stop authority, rotation, and support; prohibited individual surveillance uses | Aggregate task concurrency, alerts, after-hours load, protected recovery, perceived control, absence and turnover, fatigue-sensitive error | Does scaling depend on hidden overload, and are organizational controls acting before individual prompts? |

| Domain              | Required definition                                                                                  | Primary measures or evidence                                                                   | Release question                                                          |
|---------------------|------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| Retained capability | Independent specification, verification, challenge, exception handling, apprenticeship, and takeover | Periodic unaided tests; seeded-error detection; calibration; recovery performance; progression | Who can recognize and manage failure after routine practice is automated? |
| Decision            | Approved authority envelope, residual exposure, evidence owner, next review trigger                  | Signed release decision; exceptions; monitoring thresholds; rollback trigger                   | What evidence would contract, expand, or stop the deployment?             |

## C Case evidence matrix

**Table 12:** Claim-to-evidence matrix for principal studies and cases

| Source; status                                                        | Direct observation                                                                                                                                      | Boundary interpretation                                                                                  | Where inference stops                                                                                       |
|-----------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| <a href="#">Brynjolfsson et al. [2025]</a> ; peer reviewed            | 5,172 support agents; 15.2% preferred estimate for issues resolved per hour, concentrated among less-experienced workers                                | Frequent labels, mature knowledge, and worker edit/reject create an acceptance-rich augmentation setting | One company and nonrandom rollout; not full autonomy, profit, or general service work                       |
| <a href="#">Cui et al. [2026]</a> ; peer reviewed                     | Three randomized field experiments, 4,867 developers; pooled IV/LATE for induced use of 26.08% more completed tasks, SE 10.3%                           | Causal task-level gain for compilers can occur in ordinary enterprise settings                           | Completed work is not release quality, incidents, maintainability, or firm value; site estimates were noisy |
| <a href="#">Dell’Acqua et al. [2026b]</a> ; peer reviewed             | 758 consultants; faster, more, and higher-quality work inside tested frontier; lower correctness on selected outside task                               | Task routing and local evaluation determine whether assistance raises value                              | Constructed tasks, dated model, one firm; outside task deliberately selected                                |
| <a href="#">Dell’Acqua et al. [2026a]</a> ; peer reviewed             | 791 P&G professionals; individuals with AI matched teams without AI; selection still benefited from people                                              | Generation and evaluative selection are separable stages                                                 | Short innovation exercise; not launched-product success or long-run team design                             |
| <a href="#">Demirer et al. [2026]</a> ; working paper                 | Over 100,000 GitHub developers; effects attenuated from commits to projects and releases; no total app-usage increase                                   | Consistent with complementarity and downstream weak links                                                | Nonrandom use, changing tools, digital work; task composition and demand are alternatives                   |
| <a href="#">He et al. [2026]</a> ; preprint                           | 802 developers, 196,212 PRs; 2.09x per-capita merged PRs, roughly doubled reviewer load, more automated review                                          | Candidate and review systems adapted together under high AI intensity                                    | One young AI-forward company; nonrandom; merge/revert are coarse quality proxies                            |
| <a href="#">Becker et al. [2025]</a> ; preprint                       | 16 experienced maintainers, 246 mature-repository issues; 19% slower with AI while believing faster                                                     | Context acquisition, review, and calibration can reverse perceived speed                                 | Small selected sample, early-2025 tools, familiar mature repositories                                       |
| <a href="#">Frommgen et al. [2024]</a> ; peer reviewed                | Google system applied ML edits addressing 7.5% of reviewer comments; authors estimated large time savings                                               | Narrow, human-initiated, locally verifiable task inside an existing gate                                 | Author-estimated hours; no independent downstream-defect or economic estimate                               |
| <a href="#">Government Digital Service [2025]</a> ; government report | 20,000 licenses across 12 organizations; telemetry for about 14,500 users and 7,115 survey responses in a three-month uncontrolled trial                | Permissions, content, training, support, and oversight are deployment complements                        | Self-reported time; selected response and telemetry populations; no audited organization-level value        |
| <a href="#">Klarna Group plc [2026]</a> ; filing                      | 31 million cumulative conversations since launch; 80% of 2025 service chats; separately dated time, repeat, and savings measures; human option retained | Bounded high-volume lane plus human escape path                                                          | Mixed periods, internal methods, nonrandomized; filing does not prove all quality claims                    |

| Source; status                                                                                               | Direct observation                                                                                                                                                                                                                               | Boundary interpretation                                                                                                                                | Where inference stops                                                                                                                                                             |
|--------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| OpenAI [2025]; vendor case                                                                                   | Vendor reports more than 98% of advisor teams actively using a knowledge assistant; controlled corpus and regression evaluation                                                                                                                  | Grounded research with advisor acceptance, not autonomous fiduciary judgment                                                                           | Vendor-authored; no controlled client, compliance, or revenue outcome                                                                                                             |
| BBVA [2025]; company release                                                                                 | Expansion to 11,000 licenses; reporting users said they saved 2.8 hours per week, respondent count undisclosed                                                                                                                                   | Enterprise access accompanied by security, compliance, learning, and internal community                                                                | Self-report; no audited redeployment or accepted-value measure                                                                                                                    |
| Duolingo, Inc. [2025, 2026]; company release and filing<br>Kirkos [2023]; company case                       | 148 courses launched in under a year and later increase in published course units<br>Large proprietary company/person dataset used for search, normalization, and similarity; human outreach and judgment retained                               | Shared-content pipeline scales generation while experts validate release<br>Observed deployment boundary lies before relational and capital commitment | AI confounded with tooling; review labor and learning efficacy undisclosed<br>No counterfactual, false-negative denominator, cost ledger, optimum, or return attribution          |
| Bonelli [2026]; peer reviewed                                                                                | Data-driven VCs expanded deals/AUM and the survival-and-follow-on composite but selected fewer rare major successes and less novelty                                                                                                             | Scaling prediction can shift attention toward legible historical patterns                                                                              | Adoption proxy and observational limits; continuous-treatment DiD assumptions; return result insignificant                                                                        |
| Cowgill et al. [2026]; peer reviewed                                                                         | Sender access to ChatGPT reduced average screening accuracy and increased demand for corroborating signals                                                                                                                                       | Cheap polish can reduce stylistic signal value and move work into diligence                                                                            | Access is not artifact-level use; context-specific experiments and heterogeneous sender effects                                                                                   |
| Replit [2025]; company incident response                                                                     | Agent deleted production-connected database; rollback restored it; dev/prod and action controls were strengthened                                                                                                                                | Environment separation, permissions, and rollback bound an autonomous failure                                                                          | One self-reported case; no incidence rate or precise counterfactual                                                                                                               |
| CoreWeave, Inc. [2025], Cerebras Systems Inc. [2024]; filings<br>Google Cloud KORA [2025]; industry research | High revenue concentration in one or two customers during disclosed periods<br>4,867 professionals; a composite AI-adoption score was positively associated with throughput and self-assessed product performance, and with delivery instability | Capital and customer scale can create common-mode commercial dependency<br>Platform, testing, feedback, and stability are complements to generation    | Two infrastructure firms and historical periods; not sector prevalence or a failure forecast<br>Observational, self-reported, and software-specific; association is not causation |

## References

- Gene M. Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18–20, 1967, Spring Joint Computer Conference*, pages 483–485, 1967. doi: 10.1145/1465482.1465560.
- Anthropic. Measuring AI agent autonomy in practice. Vendor research report, February 2026. URL <https://www.anthropic.com/research/measuring-agent-autonomy>.
- Leopold Aschenbrenner. Situational awareness: The decade ahead. Scenario essay, 2024. URL <https://situational-awareness.ai/>. Accessed August 21, 2026.
- Chloe Autio, Reva Schwartz, Jesse Duniety, Shomik Jain, Martin Stanley, Elham Tabassi, Patrick Hall, and Kamie Roberts. Artificial intelligence risk management framework: Generative artificial intelligence profile. Technical Report NIST AI 600-1, National Institute of Standards and Technology, 2024.
- Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, 1983. doi: 10.1016/0005-1098(83)90046-8.

- Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabakci, and Rei Mariman. Generative ai without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, 2025. doi: 10.1073/pnas.2422633122.
- BBVA. Bbva expands its agreement with openai to 11,000 chatgpt licences for the bank’s employees. Company release, 2025. URL <https://www.bbva.com/en/innovation/bbva-expands-its-agreement-with-openai-to-11000-chatgpt-licences-for-the-banks-employees/>.
- Joel Becker, Nate Rush, Beth Barnes, and David Rein. Measuring the impact of early-2025 ai on experienced open-source developer productivity. *arXiv preprint arXiv:2507.09089*, 2025. doi: 10.48550/arXiv.2507.09089.
- Eshta Bhardwaj, Rohan Alexander, and Christoph Becker. Limits to AI growth: The ecological and social consequences of scaling. *arXiv preprint arXiv:2501.17980*, 2025. doi: 10.48550/arXiv.2501.17980. URL <https://arxiv.org/abs/2501.17980>.
- Maxime Bonelli. Data-driven investors. *The Review of Financial Studies*, 39(7):1909–1969, July 2026. doi: 10.1093/rfs/hhaf078.
- Timothy F. Bresnahan, Erik Brynjolfsson, and Lorin M. Hitt. Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *The Quarterly Journal of Economics*, 117(1):339–376, February 2002. doi: 10.1162/003355302753399526.
- Erik Brynjolfsson, Daniel Rock, and Chad Syverson. The productivity j-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1):333–372, 2021. doi: 10.1257/mac.20180386.
- Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. Generative ai at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044.
- Krzysztof Budzyń, Marcin Romańczyk, Diana Kitale, Paweł Kołodziej, Marek Bugajski, Hans O. Adami, Johannes Blom, Marek Buszkiewicz, Natalie Halvorsen, Cesare Hassan, Tomasz Romańczyk, Øyvind Holme, Krzysztof Jarus, Shona Fielding, Melina Kunar, Maria Pellise, Nastazja Pilonis, Michał Filip Kamiński, Mette Kalager, Michael Bretthauer, and Yuichi Mori. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10):896–903, October 2025. doi: 10.1016/S2468-1253(25)00133-5.
- Heejae Byun and George Gerchow. Securing ai agent toolchains: Exploiting and hardening mcp servers. RSAC 2026 conference presentation, 2026. URL <https://www.rsaconference.com/library/presentation/usa/2026/securing-ai-agent-toolchains-exploiting-and-hardening-mcp-servers>.
- Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei A. Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems*, volume 38, pages 135676–135729. Curran Associates, Inc., 2025. doi: 10.52202/085713-4082. URL [https://proceedings.neurips.cc/paper\\_files/paper/2025/hash/b1041e52d3be19f0a9bc491657488e4a-Abs tract-Datasets\\_and\\_Benchmarks\\_Track.html](https://proceedings.neurips.cc/paper_files/paper/2025/hash/b1041e52d3be19f0a9bc491657488e4a-Abs tract-Datasets_and_Benchmarks_Track.html).
- Cerebras Systems Inc. Registration statement on form s-1. U.S. Securities and Exchange Commission filing, September 2024. URL <https://www.sec.gov/Archives/edgar/data/2021728/000162828024041596/cerebras-sx1.htm>. Filed September 30, 2024; SEC accession no. 0001628280-24-041596.
- Aaron Chatterji, David Holtz, Neel Rakholia, Prasanna Tambe, and Gawesha Weeratunga. How organizations use ai: Evidence from chatgpt. *arXiv preprint arXiv:2608.12236*, 2026. URL <https://arxiv.org/abs/2608.12236>.
- Przemysław Chojacki. An operational kardashev-style scale for autonomous AI – towards AGI and superintelligence. *arXiv preprint arXiv:2511.13411*, 2025. doi: 10.48550/arXiv.2511.13411. URL <https://arxiv.org/abs/2511.13411>.

- CoreWeave, Inc. Prospectus (form 424b4). U.S. Securities and Exchange Commission filing, March 2025. URL <https://www.sec.gov/Archives/edgar/data/1769628/000119312525067651/d899798d424b4.htm>. Prospectus dated March 27, 2025; filed March 31, 2025; SEC accession no. 0001193125-25-067651.
- Bo Cowgill, Pablo Hernandez-Lagos, and Nataliya Langburd Wright. Does ai cheapen talk? theory and evidence from global entrepreneurship and hiring. *Management Science*, 2026. doi: 10.1287/mnsc.2024.07027.
- Kevin Zheyuan Cui, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz. The effects of generative ai on high-skilled work: Evidence from three field experiments with software developers. *Management Science*, 2026. doi: 10.1287/mnsc.2025.00535.
- Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramer. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2636.
- DEF CON. Invitation is all you need: Hacking gemini. DEF CON 33 session listing, 2025. URL [https://defcon.org/html/defcon-33/dc-33-speakers.html#content\\_60375](https://defcon.org/html/defcon-33/dc-33-speakers.html#content_60375).
- Fabrizio Dell’Acqua, Charles Ayoubi, Hila Lifshitz, Raffaella Sadun, Ethan Mollick, Lilach Mollick, Yi Han, Jeff Goldman, Hari Nair, Stew Taub, and Karim R. Lakhani. The cybernetic teammate: A field experiment on generative ai and teamwork. *Organization Science*, 37(4):1217–1242, 2026a. doi: 10.1287/orsc.2025.20702.
- Fabrizio Dell’Acqua, Edward McFowland, Ethan Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Kraye, Francois Candelon, and Karim R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2):403–423, 2026b. doi: 10.1287/orsc.2025.21838.
- Mert Demirer, Andrey Fradkin, Nadav Tadelis, and Sida Peng. The emerging market for intelligence: Pricing, supply, and demand for llms. Technical Report 34608, National Bureau of Economic Research, 2025.
- Mert Demirer, Leon Musolff, and Liyuan Yang. Writing code vs. shipping code: Productivity effects across generations of ai coding tools. Technical Report 35275, National Bureau of Economic Research, 2026.
- Duolingo, Inc. Duolingo launches 148 new language courses. Company release dated April 30, 2025, April 2025. URL <https://investors.duolingo.com/news-releases/news-release-details/duolingo-launches-148-new-language-courses>.
- Duolingo, Inc. Q1 2026 shareholder letter. Exhibit 99.2 to Form 8-K, filed May 4, 2026, May 2026. URL <https://www.sec.gov/Archives/edgar/data/1562088/000162828026029790/q1fy26duolingo3-31x26share.htm>. SEC accession no. 0001628280-26-029790.
- Nelly Dux, Cristina Alaimo, Philippe Roussiere, and Abhishek Kumar Mishra. Governance by design: Architecting agentic ai for organizational learning and scalable autonomy. *arXiv preprint arXiv:2605.20210*, 2026. doi: 10.48550/arXiv.2605.20210.
- Mica R. Endsley and Esin O. Kiris. The out-of-the-loop performance problem and level of control in automation. *Human Factors*, 37(2):381–394, June 1995. doi: 10.1518/001872095779064555.
- Eurostat. 20% of eu enterprises use ai technologies. Eurostat news article dated December 11, 2025, December 2025. URL <https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20251211-2>.
- FinOps Foundation. State of FinOps 2026. Community survey report, 2026. URL <https://data.finops.org/>. 1,192 total respondents; AI-spend item N=693; skills item N=695.
- Alexander Frommgen, Jacob Austin, Peter Choy, Nimesh Ghelani, Lera Kharatyan, Gabriela Surita, Elena Khrapko, Pascal Lamblin, Pierre-Antoine Manzagol, Marcus Revaj, et al. Resolving code review comments with machine learning. In *ICSE-SEIP ’24: Proceedings of the 46th IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice*, pages 204–215. Association for Computing Machinery, 2024. doi: 10.1145/3639477.3639746. URL <https://research.google/pubs/resolving-code-review-comments-with-machine-learning/>.

- Claudine Gartenberg, Sharique Hasan, Alex Murray, and Lamar Pierce. More versus better: Artificial intelligence, incentives, and the emerging crisis in peer review. *Organization Science*, 37(3):795–812, 2026. doi: 10.1287/orsc.2026.ed.v37.n3.
- Tim Genewein, Matija Franklin, Alexander Lerchner, Laurent Orseau, Samuel Albanie, Adam Bales, Cole Wyeth, Stephanie Chan, Iason Gabriel, Joel Z. Leibo, Allan Dafoe, Marcus Hutter, Thore Graepel, and Shane Legg. From AGI to ASI. *arXiv preprint arXiv:2606.12683*, 2026. doi: 10.48550/arXiv.2606.12683. URL <https://arxiv.org/abs/2606.12683>.
- Google Cloud. Ai and ml perspective: Reliability. Google Cloud Architecture Framework, 2025a. URL <https://docs.cloud.google.com/architecture/framework/perspectives/ai-ml/reliability>.
- Google Cloud. June 12, 2025 service disruption: Incident report. Official service-health incident report, June 2025b. URL <https://status.cloud.google.com/incidents/ow5i3PPK96RduMcb1SsW>.
- Google Cloud DORA. State of ai-assisted software development 2025. Technical report, Google Cloud, 2025. URL <https://cloud.google.com/resources/content/2025-dora-ai-assisted-software-development-report>.
- Google DeepMind. Frontier safety framework, version 3.1. Technical framework, 2026. URL <https://deepmind.google/frontier-safety/>.
- Government Digital Service. Microsoft 365 copilot experiment: Cross-government findings report. Technical report, Government Digital Service, London, June 2025. URL <https://www.gov.uk/government/publications/microsoft-365-copilot-experiment-cross-government-findings-report>.
- Adam Grundy, Cory Breau, and Dhanapati Khatiwoda. Large firms with at least 20 employees biggest ai users. U.S. Census Bureau, America Counts, May 2026. URL <https://www.census.gov/library/stories/2026/05/ai-use-businesses.html>.
- Hao He, Shyam Agarwal, Yegor Denisov-Blanch, Pavel Azaletskiy, Sanmi Koyejo, and Bogdan Vasilescu. Ai writes faster than humans can review: A longitudinal study of an enterprise 2x mandate. *arXiv preprint arXiv:2607.01904*, 2026. URL <https://arxiv.org/abs/2607.01904>.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030, 2022. URL <https://arxiv.org/abs/2203.15556>.
- Muhammad Zia Hydari and Farooq Muzaffar. Redrawing the AI map: A theory of accountability boundaries in agentic ecosystems. *arXiv preprint arXiv:2605.23179*, 2026. doi: 10.48550/arXiv.2605.23179. URL <https://arxiv.org/abs/2605.23179>.
- Wei Jiang, Junyoung Park, Rachel Jiqui Xiao, and Shen Zhang. AI and the extended workday: Productivity, contracting efficiency, and distribution of rents. Technical Report 33536, National Bureau of Economic Research, February 2025.
- Drew Johnston, David Holtz, Alex Martin Richmond, Christopher Ong, Prasanna Tambe, and Aaron Chatterji. The shift to agentic AI: Evidence from Codex, June 2026. URL <https://arxiv.org/abs/2606.26959>.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL <https://arxiv.org/abs/2001.08361>.
- Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research*, 2025. URL <https://mlanthology.org/tmlr/2025/kapoor2025tmlr-ai/>.

- Tahmina Karimova. AI systems at work: A changing psychosocial work environment. ILO Working Paper 170, International Labour Organization, April 2026. URL <https://www.ilo.org/publications/ai-systems-work-changing-psychosocial-work-environment>.
- Hyunjin Kim and Rembrand Koning. Ai-native firms. Technical Report 26-090, Harvard Business School, 2026. URL <https://www.insead.edu/faculty-research/publications/working-papers/ai-native-firms>.
- Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Towards a science of scaling agent systems. *arXiv preprint arXiv:2512.08296*, 2025. doi: 10.48550/arXiv.2512.08296. URL <https://arxiv.org/abs/2512.08296>. Version 3, revised April 2026.
- J. F. C. Kingman. The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society*, 57(4):902–904, October 1961. doi: 10.1017/S0305004100036094.
- Ilya Kirnos. Vc gpt: How llms are strengthening signalfire’s in-house ai. SignalFire, 2023. URL <https://www.signalfire.com/blog/signalfire-beacon-ai>.
- Klarna Group plc. Annual report for the year ended december 31, 2025 (form 20-f). U.S. Securities and Exchange Commission filing, 2026. URL <https://www.sec.gov/Archives/edgar/data/2003292/000200329226000007/klar-20251231.htm>.
- Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean. Ai 2027. Scenario forecast, 2025. URL <https://ai-2027.com/>. Accessed August 21, 2026.
- Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Megan Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sebastian von Arx, et al. Measuring ai ability to complete long software tasks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3086.
- Marguerita Lane, Morgan Williams, and Stijn Broecke. The impact of AI on the workplace: Main findings from the OECD AI surveys of employers and workers. OECD Social, Employment and Migration Working Papers 288, OECD Publishing, March 2023.
- Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025. doi: 10.1145/3706598.3713778.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL <https://arxiv.org/abs/2211.09110>.
- John D. C. Little. A proof for the queuing formula:  $L = \lambda W$ . *Operations Research*, 9(3):383–387, June 1961. doi: 10.1287/opre.9.3.383.
- Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathaniel Khazam, Jad Kabbara, et al. A large-scale audit of dataset licensing and attribution in ai. *Nature Machine Intelligence*, 6:975–987, 2024. doi: 10.1038/s42256-024-00878-8.
- David Lyell and Enrico Coiera. Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association*, 24(2):423–431, 2017. doi: 10.1093/jamia/ocw105.
- Brendan Lynch and Richard Harang. From prompts to pwns: Exploiting and securing ai agents. Black Hat USA conference presentation, 2025. URL <https://i.blackhat.com/BH-USA-25/Presentations/US-25-Lynch-From-Prompts-to-Pwns.pdf>.

- Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Position: Levels of agi for operationalizing progress on the path to agi. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 36308–36321. PMLR, 2024. URL <https://proceedings.mlr.press/v235/morris24b.html>.
- National Cybersecurity Center of Excellence. Accelerating the adoption of software and ai agent identity and authorization. Draft concept paper, National Institute of Standards and Technology, February 2026. URL <https://www.nccoe.nist.gov/publications/other/accelerating-adoption-software-and-ai-agent-identity-and-authorization-concept>.
- OECD. Venture capital investments in artificial intelligence through 2025. OECD Policy Briefs 50, OECD Publishing, Paris, February 2026.
- OpenAI. Morgan stanley uses ai evals to shape the future of financial services. Vendor case study, 2025. URL <https://openai.com/index/morgan-stanley/>. Accessed August 21, 2026.
- OpenAI. Bbva puts ai at the core of banking with openai. Vendor case study, 2026a. URL <https://openai.com/index/bbva/>.
- OpenAI. June 2, 2026 service disruption: Incident write-up. Official status incident report, June 2026b. URL <https://status.openai.com/incidents/01KT5XJ5ATD6RMYP908WS69FVD/write-up>.
- OpenTelemetry. Generative ai semantic conventions. OpenTelemetry specification, 2026. URL <https://opentelemetry.io/docs/specs/semconv/registry/attributes/gen-ai/>. Accessed August 21, 2026; development-status conventions.
- OWASP GenAI Security Project. Owasp top 10 for agentic applications. Community security guidance, 2025. URL <https://genai.owasp.org/2025/12/09/owasp-top-10-for-agentic-applications-the-benchmark-for-agentic-security-in-the-age-of-autonomous-ai/>.
- Raja Parasuraman and Dietrich H. Manzey. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410, 2010. doi: 10.1177/0018720810376055.
- Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297, May 2000. doi: 10.1109/3468.844354.
- PitchBook and National Venture Capital Association. Pitchbook-nvca venture monitor: Q2 2026. Technical report, National Venture Capital Association, 2026. URL <https://nvca.org/wp-content/uploads/2026/07/Q2-2026-PitchBook-NVCA-Venture-Monitor.pdf>.
- PyTorch. torch.load with weights\_only=true remote code execution (ghsa-53q9-r3pm-6pq6). Security advisory, 2025. URL <https://github.com/pytorch/pytorch/security/advisories/GHSA-53q9-r3pm-6pq6>.
- Anita Rao, Andrew Keller, Neha Kalra, Ryan Steed, Kweku Kwegyir-Aggrey, Kevin Klyman, Diane Staheli, and Amanda Bergman. Challenges to the monitoring of deployed AI systems. Technical Report NIST AI 800-4, National Institute of Standards and Technology, March 2026. URL <https://doi.org/10.6028/NIST.AI.800-4>.
- Replit. Doubling down on our commitment to secure vibe coding. Company incident response, July 2025. URL <https://replit.com/blog/doubling-down-on-our-commitment-to-secure-vibe-coding>.
- James Ryseff, Brandon F. De Bruhl, and Sydne J. Newberry. The root causes of failure for artificial intelligence projects and how they can succeed: Avoiding the anti-patterns of ai. Technical Report RR-A2680-1, RAND Corporation, Santa Monica, CA, August 2024.
- J. H. Shen, D. Zhu, S. Srinivasan, H. Sleight, L. T. Wagner III, M. J. Matthews, E. Jones, and Jascha Sohl-Dickstein. Ai organizations are more effective but less aligned than individual agents. *arXiv preprint arXiv:2604.10290*, 2026. doi: 10.48550/arXiv.2604.10290.

- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. URL <https://arxiv.org/abs/2408.03314>.
- Stanford Institute for Human-Centered Artificial Intelligence. The 2026 ai index report. Technical report, Stanford University, 2026. URL <https://hai.stanford.edu/ai-index/2026-ai-index-report>.
- UK National Cyber Security Centre. Prompt injection is not sql injection. Technical blog, December 2025. URL <https://www.ncsc.gov.uk/blog-post/prompt-injection-is-not-sql-injection>.
- U.S. Securities and Exchange Commission. Sec charges two investment advisers with making false and misleading statements about their use of artificial intelligence. Press Release 2024-36, 2024. URL <https://www.sec.gov/newsroom/press-releases/2024-36>.
- U.S. Securities and Exchange Commission, Division of Corporation Finance. Comment letter regarding klarna group plc, amendment no. 1 to registration statement on form f-1, file no. 333-285826. SEC correspondence dated May 30, 2025, May 2025. URL <https://www.sec.gov/Archives/edgar/data/2003292/000000000025005760/filename1.pdf>.
- World Health Organization. Guidelines on mental health at work. Technical report, World Health Organization, 2022. URL <https://www.who.int/publications/i/item/9789240053052>.
- Yuxing Xiang, Xue Li, Kun Qian, Yan Zhang, Wenyuan Yu, Ennan Zhai, Xin Jin, and Jingren Zhou. ServeGen: Workload characterization and generation of large language model serving in production. In *23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI 26)*, pages 1845–1859, Renton, WA, May 2026. USENIX Association. ISBN 978-1-939133-54-0. URL <https://www.usenix.org/conference/nsdi26/presentation/xiang-servegen>.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. Tau-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=roNSXZpUDN>.