The Implementable Scale of Agentic Automation:
A Unified Theory of Verification, Isolation, and the Stages of AI-Native Software Engineering

Alexander Quispe*

July 2026

Abstract

We develop a unified economic model of agentic coding automation. An organization chooses how many AI agents to run, how intensively agents verify their own output, and what share of surviving output receives independent verification, subject to human review capacity, filesystem and runtime isolation capital, and finite useful demand for validated changes. Verification is modeled as a layered acceptance technology with a correlated-error floor: an agent that misreads a specification can write both the wrong code and the test that approves it, so self-verification alone can never drive false acceptance below a strictly positive bound; only independent verification can. The implementable agent scale is the minimum of an unconstrained profit-maximizing scale, a verification-capacity ceiling, an isolation ceiling, and a useful-demand ceiling, and we state exactly the assumptions under which this minimum formula is a theorem rather than an accounting identity. The five stages of the practitioner maturity ladder (gated, assisted, parallel, supervised autonomy, AI-native) emerge as the identity of the binding constraint, and stage transitions are discounted investments that break the currently binding bottleneck. The model delivers closed forms for the optimal delegation share, the optimal self-verification intensity, the verification ceiling, and the interior agent scale, together with testable predictions that separate usage from return.

Keywords: agentic AI, automation stages, verification, delegation, software engineering.
JEL codes: D24, J24, L23, O33.

Contents

1 Introduction **4**

2 Related work **4**

3 Environment and primitives **5**
3.1 Agents, tasks, and time 5
3.2 Humans 5
3.3 Agents and capital 5
3.4 The capital stocks in plain language 6
3.5 Verification primitives 7
3.6 Flows and costs 8

*Pontificia Universidad Católica del Perú. This paper unifies and replaces an earlier tutorial note on the Cherny 0–4 maturity ladder. All derivations are self-contained; every algebraic step is reproduced in the appendix so results can be verified by hand.

4 The task-level delegation model **8**
4.1 Certainty equivalents . . . 8
4.2 The manual mode . . . 9
4.3 The delegated mode . . . 9
4.4 Optimal delegation . . . 9
4.5 Thresholds and the stage-1 menu . . . 10
4.6 Menu nesting and its limits . . . 11
4.7 From task activation to the demand for validated changes . . . 11

5 The verification technology **11**
5.1 Layered acceptance . . . 11
5.2 Guardrail reliability . . . 12

6 The organization's problem and the implementable scale **12**
6.1 Review standards and flows . . . 12
6.2 The objective . . . 13
6.3 The implementable scale . . . 13
6.4 Tokens are not enough . . . 14

7 Optimal verification policy **14**
7.1 Self-verification intensity . . . 14
7.2 Coverage, substitutes, and complements . . . 15

8 Soft utilization: risk rises with scale **15**

9 Stages as binding constraints **16**

10 Stage transitions **16**

11 Empirical content **17**
11.1 Mapping model objects to telemetry . . . 17
11.2 Predictions . . . 18

12 Conclusion **18**

A The CARA–Normal certainty equivalent **19**

B Proofs for Section 4 **19**
B.1 Proposition 1 (optimal delegation) . . . 19
B.2 Envelope value (13) . . . 20
B.3 Corollary 1 (perfect-square adoption condition) . . . 20
B.4 Proposition 2 (menu nesting) . . . 20

C Proof of Lemma 2 (demand bridge) **20**

D Proofs for Section 6 **21**
D.1 Theorem 1 . . . 21
D.2 Corollary 2 . . . 21
D.3 Proposition 4 (interior scale, linear demand) . . . 21
D.4 Proposition 5 (bottleneck investment) . . . 21

E Proof for Section 7 **22**
E.1 Proposition 6 . . . 22
E.2 Proposition 7 . . . 22

F Proofs for Section 8 **22**
F.1 Lemma 4 . . . 22
F.2 Proposition 8 . . . 22

G Proof of Proposition 9 **23**

H Soft isolation: microfoundation **23**

I Dynamic extensions **24**
I.1 Guardrail reliability . . . 24
I.2 Codification and the diffusion gap . . . 24
I.3 Regulatory and legacy burden . . . 24
I.4 The ragged frontier . . . 24

J Notation **24**

1 Introduction

Coding agents have moved from autocomplete assistants to systems that plan, edit, test, and open pull requests with limited human involvement. Practitioners describe this evolution as a maturity ladder: Stage 0 (*gated*: the organization blocks agent use), Stage 1 (*assisted*: one supervised agent), Stage 2 (*parallel*: roughly ten agents orchestrated by one engineer), Stage 3 (*supervised autonomy*: roughly one hundred agents running routines), and Stage 4 (*AI-native*: thousands of agents steered by intent and monitored by exception) [1, 2]. The recurring practitioner observation is that tokens alone do not move a team up this ladder: each stage has a different binding constraint, and advancing requires breaking that specific bottleneck.

This paper turns that observation into a theorem. We build one model in which a developer or organization chooses three things — the number of agents n, the intensity of agent *self-verification* x, and the coverage of *independent verification* g — given four stocks of complementary capital (guardrail capital K^G, process-codification capital K^C, filesystem-isolation capital K^F, runtime-isolation capital K^R), human review capacity H^V, and a finite pool of useful tasks. The central objects are:

(i) a *task-level delegation model* that determines when a task is worth doing at all and how much of its execution is delegated to an agent (Section 4);

(ii) a *layered verification technology* in which candidates pass the agent's own checks and then, with some coverage probability, independent checks; the technology has a correlated-error floor that self-verification cannot cross (Section 5);

(iii) an *organization-level problem* whose solution is the implementable agent scale

$$n^* = \min\{n^u,\ \bar{n}^V,\ \bar{n}^I,\ \bar{n}^D\}, \tag{1}$$

the minimum of an unconstrained profit-maximizing scale, a verification-capacity ceiling, an isolation ceiling, and a useful-demand ceiling (Section 6);

(iv) a definition of *stage* as the identity of the binding constraint in (1), so that the practitioner ladder becomes bottleneck migration (Section 9);

(v) a *transition rule* in which moving up a stage is a discounted investment in the currently binding bottleneck (Section 10).

Three design choices distinguish the model from a mechanical translation of the practitioner discussion. First, the demand side is not a free parameter: the task-level activation model of Section 4 *is* the demand curve for validated changes used in the organization-level problem, through an exact change-of-variables identity (Lemma 2). Second, verification risk is not mean-zero noise: a falsely accepted change produces a strictly negative payoff (an incident), and the probability of false acceptance depends on who verifies, how carefully, and with what tooling. Third, every constraint that appears in the model also binds somewhere: each ceiling in (1) is derived, imposed, and priced, and we state precisely when the minimum formula is valid (Theorem 1) and when one of its arms is redundant (Corollary 2).

2 Related work

The human-automation literature classifies levels of automation and the functions automated [7, 8]. Task-based models of automation in economics study which tasks machines absorb and which new

tasks technologies create [9]. Our micro block borrows the delegation logic of authority models [10]: the human retains specification and verification authority while transferring execution. The exception-based review structure of high stages is a knowledge hierarchy in the sense of [11]: routine matters are handled by the lower layer (agents plus automated guardrails), and only exceptions are referred to scarce human attention. The stage-transition rule is a standard discounted adoption condition; the option-value qualification follows [12]. Engineering documentation describes multi-agent orchestration, hooks, and isolation tooling [5, 6, 4], and observed usage patterns show humans retaining planning and verification decisions while delegating execution [3]. What appears new here is a single maximization problem in which agent count, self-verification, and independent verification are jointly chosen against verification, isolation, and demand constraints, and in which the practitioner stage ladder is the path of the binding constraint.

3 Environment and primitives

3.1 Agents, tasks, and time

A developer or engineering team is indexed by i; discrete periods are $t = 0, 1, 2, \ldots$. There is a pool of potential tasks j (bug fixes, features, migrations, dependency upgrades, documentation, maintenance). Task j has an *opportunity value* $\omega_{ij,t} \in \mathbb{R}$, drawn from a continuous, strictly increasing distribution F on an interval support, and an *activation cost* $b_{ij,t} \geq 0$. The mass of potential tasks per period is J.

3.2 Humans

The human has task-specific execution skill $s_{ij,t} \in [0, 1]$ and general specification-and-verification ability $a_i \geq 0$: the ability to state objectives, decompose work, design tests, and judge output. The human is uncertain about her own productivity match θ_{ij} on task j:

$$\theta_{ij} \mid \mathcal{F}_{it} \sim N\left(\mu_{ij,t},\, \frac{1}{\tau_{ij,t}}\right), \tag{2}$$

where $\mu_{ij,t}$ is the posterior mean and $\tau_{ij,t}$ the posterior *precision*. Preferences over risky payoffs are CARA with coefficient $\rho > 0$.

3.3 Agents and capital

Model capability is A; agent execution competence in the task-level model is $z(A)$ with $z' > 0$. The organization holds four capital stocks, rented at per-period rates r_G, r_C, r_F, r_R:

Stock	Content
K^G (guardrail capital)	tests, CI gates, security scanners, policy-as-code, hooks
K^C (codification capital)	shared prompts, rules files, evals, documented workflows
K^F (filesystem isolation)	worktrees, clones, branch discipline
K^R (runtime isolation)	containers, namespaces, per-agent databases, preview environments

Gating friction $\gamma_\ell \geq 0$ is the per-task cost of *accessing* the agentic mode at stage ℓ (approvals, procurement, compliance path). Friction gates the AI mode only; it never taxes manual work. Guardrail capital and gating friction are distinct objects with opposite welfare signs, and they never share a symbol.

3.4 The capital stocks in plain language

The table above states the four stocks in software terms, but nothing in the economics depends on those terms. This subsection restates them for readers outside software engineering, with one running analogy: deploying coding agents is like staffing a restaurant with a brigade of talented, brand-new cooks. Model capability A is the cooks' raw talent, hired from outside at a market price; the four capital stocks are everything the restaurant itself must build before talent turns into meals served. Figure 1 places all four stocks in the production line, and Figure 2 pictures the scale formula they feed into (Theorem 1). Readers fluent in the software objects can skip to Section 5 without loss.

Definition 1 (Guardrail capital K^G)**.** *Automated machinery that checks work before any human looks at it.* In software: automated test suites, continuous-integration gates (checks that every proposed change must pass before it can enter the shared codebase), security scanners, and policy rules enforced by machine rather than by memo. Kitchen analogue: thermometers, timers, and smoke alarms. A thermometer catches undercooked chicken without the head chef tasting every plate. In the model, guardrail capital makes agent self-checking cheap (c_x is decreasing in K^G) and determines whether external review can catch the failures that actually occur (reliability R^G, Section 5).

Definition 2 (Codification capital K^C)**.** *The organization's knowledge, written down where an agent can read it.* In software: shared prompt libraries, rules files that state the house conventions, evaluation suites, documented workflows. Kitchen analogue: the recipe book and operations manual of a franchise. A new cook can produce the house dish on day one only if the recipe lives on paper rather than in the head chef's memory. In the model, codification is the only stock that directly raises production: candidates per agent $\zeta(A, K^C)$ and the probability $\pi_c(A, K^C)$ that a candidate is correct both increase in K^C.

Definition 3 (Filesystem-isolation capital K^F)**.** *A separate copy of the work-in-progress for each worker.* In software: git worktrees and repository clones (private working copies of the codebase) and the discipline of keeping each agent's changes on its own branch. Kitchen analogue: a cutting board and mise en place per cook; two cooks sharing one board ruin each other's preparation no matter how skilled each is. In the model, K^F/m_F is one arm of the concurrency cap (24).

Definition 4 (Runtime-isolation capital K^R)**.** *A separate place for each worker to try the work out.* In software: containers and namespaces (lightweight private computers-within-the-computer), a database per agent, preview environments where a change can be run without touching production. Kitchen analogue: a test kitchen per cook — own burner, own oven, own pantry — so one cook can burn a dish without disrupting anyone else's service. In the model, K^R/m_R is the other arm of (24), and it is typically the binding one: cutting boards are cheap, ovens are not.

Two vocabulary points. First, these are *capital* in the standard economic sense: durable stocks, costly to build, yielding services over many periods. Second, the rental rates r_m price *upkeep*, not construction: tests must track a moving codebase, manuals must stay current, sandboxes accrue cloud charges. Construction is priced separately by the one-time installation cost Φ of Section 10; no expenditure is counted twice.

Stock	Everyday analogue	What it buys in the model
K^G	thermometers, smoke alarms	cheap self-checks ($c_x \downarrow$); review that works ($R^G \uparrow$)
K^C	recipe book, operations manual	more candidates per agent ($\zeta \uparrow$), more of them right ($\pi_c \uparrow$)
K^F	a cutting board per cook	concurrency slots K^F/m_F
K^R	a test kitchen per cook	concurrency slots K^R/m_R; usually the scarce arm

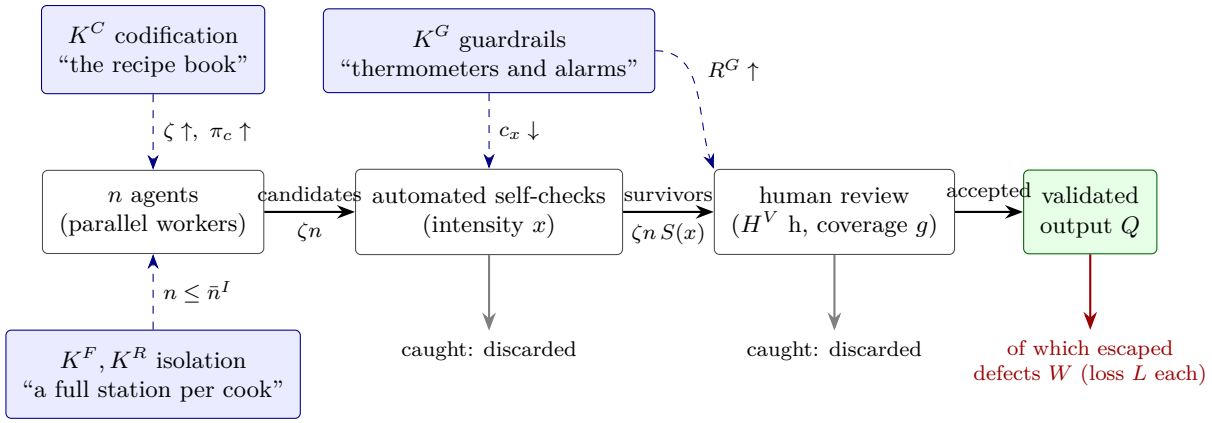

Figure 1: Where each capital stock enters the production of validated changes. Agents generate candidate changes; automated self-checks (cheap only when guardrail capital exists) filter them; human review filters the survivors. Codification capital raises the quantity and quality of candidates; isolation capital caps how many agents can work at once; the accepted flow contains the escaped defects W that both layers missed.

Example 1 (A five-agent kitchen)**.** A team rents enough compute for ten agents and asks how many it can actually run. Isolation: $K^F = 20$ workspace slots and $K^R = 8$ sandbox slots, one of each per agent ($m_F = m_R = 1$), so at most $\bar{n}^I = \min\{20, 8\} = 8$ agents fit concurrently — the ninth has a cutting board but no oven. Verification: each agent's candidates survive self-checks at rate $\zeta S = 8$ per period, every survivor is reviewed ($g = 1$) at $\hat{e} = 1$ hour, and reviewers have $H^V = 40$ hours, so $\bar{n}^V = 40/(1 \cdot 8 \cdot 1) = 5$. Demand: the backlog absorbs 45 validated changes per period and each agent delivers $\mu_Q = 5$, so $\bar{n}^D = 9$; the unconstrained profit peak is $n^u = 10$. The implementable scale is $n^* = \min\{10, 5, 8, 9\} = 5$: verification binds (Figure 2). Buying more sandboxes now is pure waste at the margin — slots six through eight already sit idle — and buying raw generation ("more tokens") is worse than useless (Proposition 5). The profitable move attacks the shortest bar: build guardrail capital so agents self-check more intensely ($x \uparrow$), cutting survivors per agent from 8 toward the floor $\zeta\pi_c = 5$ (correct candidates always survive, so review load cannot fall further), which lifts $\bar{n}^V$ from 5 toward 8 — at which point *isolation* binds and the next investment is runtime sandboxes. The shortest bar migrates; that migration is precisely the stage transition of Section 10.

3.5 Verification primitives

Each agent-produced *candidate change* is correct with probability $\pi_c(A, K^C) \in (0, 1)$, increasing in both arguments. Verification has two layers:

(a) **Self-verification** at intensity $x \ge 0$ (agent-run tests, builds, typechecks, lints): an *incorrect* candidate survives the agent's own checks with probability

$$p^{\text{SV}}(x) = \underline{p}(A) + \big(p_0 - \underline{p}(A)\big)e^{-x}, \qquad 0 < \underline{p}(A) < p_0 \le 1. \tag{3}$$

The *correlated-error floor* $\underline{p}(A) > 0$ captures the practitioner objection that the same model that misreads a specification writes both the wrong patch and the test that approves it: no amount of self-checking removes errors the agent cannot see. Correct candidates survive self-checks with probability one (self false rejection is a straightforward extension).

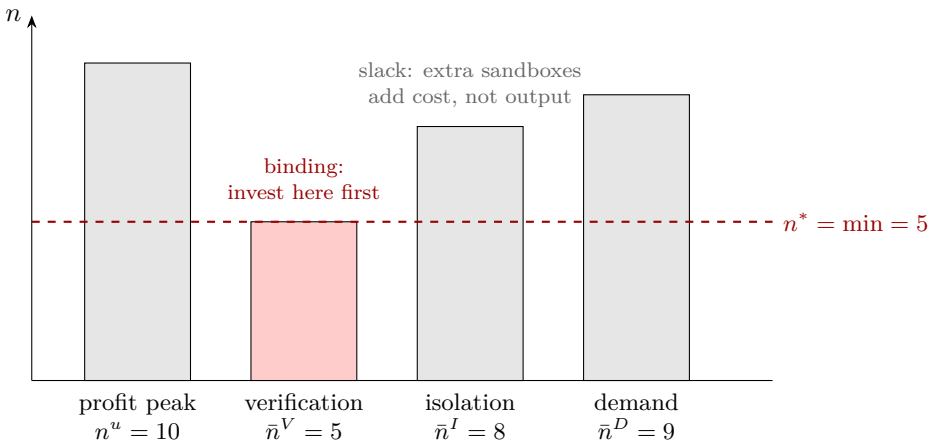

Figure 2: The implementable scale $n^* = \min\{n^u, \bar{n}^V, \bar{n}^I, \bar{n}^D\}$ (Theorem 1) with the numbers of Example 1. The shortest bar names the maturity stage; only investment that raises the shortest bar raises output, and raising a slack bar is pure waste at the margin (Proposition 5). Stage transitions are episodes in which investment makes the shortest bar migrate.

(b) **Independent verification** with coverage $g \in [0,1]$: each surviving candidate is routed, with probability g, to an external check (human review assisted by guardrails). A reviewed *incorrect* candidate is caught with probability $R^G d(e)$, where e is review effort (hours) per reviewed candidate, $d(\cdot)$ is an increasing, concave detection function with $d(0) = 0$, and $R^G \in [0,1]$ is guardrail reliability (the probability that the guardrail suite is capable of flagging the relevant failure class; its maintenance dynamics are in Appendix I). A reviewed *correct* candidate is falsely rejected with probability $q^{\mathrm{FR}}(e)$, decreasing and convex in e.

An accepted incorrect change causes an incident with loss $L > 0$.

3.6 Flows and costs

With n agents, candidates are generated at rate

$$C(n) = \zeta(A, K^C)\, n, \tag{4}$$

with ζ increasing in capability and codification capital, subject to the isolation cap of Section 6 (a congestion microfoundation with soft conflicts is in Appendix H). Compute cost is c_A per candidate plus c_x per unit of self-verification per candidate; c_x is decreasing in K^G (self-verification is cheap only when a test harness exists). Human verification capacity is H^V hours per period at wage w.

4 The task-level delegation model

This section determines which tasks are done and how much of each is delegated. Its output is the demand curve for validated changes used in Section 6.

4.1 Certainty equivalents

Lemma 1 (CARA–Normal certainty equivalent)**.** *If $Y \mid \mathcal{F} \sim N(m, v)$ and utility is $u(y) = -e^{-\rho y}$, the certainty equivalent is*

$$CE(Y) = m - \frac{\rho}{2} v. \tag{5}$$

The proof, from the Normal moment generating function, is Appendix A.

4.2 The manual mode

Executing task j manually yields random payoff $Y^{\text{man}} = \omega + s\theta$ (task subscripts suppressed). By (2) and Lemma 1, the net certainty-equivalent surplus after activation cost b is

$$V^{\text{man}} = \omega + s\mu - b - \frac{\rho}{2}\frac{s^2}{\tau}. \tag{6}$$

The task is done manually iff $V^{\text{man}} \geq 0$, i.e. iff $\omega \geq T^{\text{man}}$ with

$$T^{\text{man}} = b - s\mu + \frac{\rho s^2}{2\tau}. \tag{7}$$

Gating friction does *not* appear in (6)–(7): blocking AI access does not change the value of hand-written code. Stage 0 is the state in which the manual mode is the only member of the feasible menu.

4.3 The delegated mode

When friction γ is finite, the developer may delegate a share $\lambda \in [0, 1]$ of execution. The delegated share produces expected quality $a\, z(A)$ per unit (the human's ability to specify and judge multiplies agent competence), requires per-unit review cost κ_1, and carries per-unit expected incident loss

$$\varphi \equiv (1 - \pi_c)\, q_1^{\text{FA}}\, L, \tag{8}$$

where q_1^{FA} is the stage-1 false-acceptance probability from the verification technology of Section 5 evaluated at stage-1 settings. Using the mode costs a fixed access-plus-usage charge $\gamma + r_1$. Define *net delegated quality*

$$X \equiv a\, z(A) - \kappa_1 - \varphi. \tag{9}$$

Assumption 1 (Expected-loss pricing of incidents)**.** Incident losses enter the certainty equivalent in expectation. (Incidents are small-probability events diversified across the task portfolio; pricing their Bernoulli variance adds a term $\frac{\rho}{2}\lambda^2\Sigma_L$ to the risk penalty and changes no qualitative result.)

The delegated-mode payoff is $Y^{\text{del}}(\lambda) = \omega + (1 - \lambda)s\theta + \lambda\, az(A) - \gamma - r_1 - \lambda\kappa_1 - \varepsilon_L$ with $\mathbb{E}[\varepsilon_L] = \lambda\varphi$. Only θ is priced as risk (Assumption 1), and $\text{Var}\big((1 - \lambda)s\theta \mid \mathcal{F}\big) = (1 - \lambda)^2 s^2/\tau$, so Lemma 1 gives

$$V^{\text{del}}(\lambda) = \omega + (1 - \lambda)s\mu + \lambda X - \gamma - r_1 - b - \frac{\rho}{2}\frac{(1 - \lambda)^2 s^2}{\tau}. \tag{10}$$

4.4 Optimal delegation

Proposition 1 (Optimal delegation share)**.** *V^{del} is strictly concave in λ, and the optimal share is*

$$\lambda^* = \text{med}\left\{0,\ 1 - \frac{\tau\,(s\mu - X)}{\rho s^2},\ 1\right\}, \tag{11}$$

where med *denotes the median (clipping to $[0, 1]$). In particular:*

(i) full delegation ($\lambda^* = 1$) iff $X \geq s\mu$;

(ii) interior delegation iff $s\mu - \rho s^2/\tau < X < s\mu$*;*

(iii) no delegation iff $X \leq s\mu - \rho s^2/\tau$*.*

In the interior region, $\partial\lambda^*/\partial A > 0$*,* $\partial\lambda^*/\partial\rho > 0$*, and* $\partial\lambda^*/\partial\tau < 0$*.*

Derivation. Differentiate (10):

$$\frac{dV^{\text{del}}}{d\lambda} = -s\mu + X + \rho(1-\lambda)\frac{s^2}{\tau}, \qquad \frac{d^2V^{\text{del}}}{d\lambda^2} = -\frac{\rho s^2}{\tau} < 0. \tag{12}$$

Setting the first derivative to zero and solving for λ yields the interior expression in (11); the corner characterization evaluates (12) at $\lambda = 1$ and $\lambda = 0$. Full steps and the comparative statics are in Appendix B. □

Two implications deserve emphasis. *Delegation is a hedge*: the human's execution risk $(1-\lambda)^2 s^2/\tau$ falls with delegation while the agent's expected quality is deterministic, so more risk-averse developers delegate *more* ($\partial\lambda^*/\partial\rho > 0$). *Experience retains execution*: developers with precise knowledge of their own match (τ large) delegate less. Both are testable.

4.5 Thresholds and the stage-1 menu

Substituting λ^* into (10) (Appendix B) gives the delegated-mode value

$$V^{\text{del},*} = \begin{cases} \omega + X - \gamma - r_1 - b, & X \geq s\mu \quad (\lambda^* = 1), \\ \omega + X + \dfrac{\tau\,(s\mu - X)^2}{2\rho s^2} - \gamma - r_1 - b, & s\mu - \dfrac{\rho s^2}{\tau} < X < s\mu, \\ V^{\text{man}} - \gamma - r_1, & X \leq s\mu - \dfrac{\rho s^2}{\tau} \quad (\lambda^* = 0,\ \text{dominated}), \end{cases} \tag{13}$$

so the delegated-mode activation threshold in the interior case is

$$T^{\text{del}} = b + \gamma + r_1 - X - \frac{\tau\,(s\mu - X)^2}{2\rho s^2}. \tag{14}$$

The quadratic term is the *hedging value* of interior mixing.

Stage 1 offers the *menu* $\mathcal{M}_1 = \{\text{manual},\ \text{delegated}\}$, and the stage-1 threshold is defined on the menu:

$$T^1 = \min\{T^{\text{man}},\, T^{\text{del}}\} \leq T^{\text{man}} = T^0. \tag{15}$$

Because the manual mode is a member of every stage's menu, the inequality holds *by construction*, not by assumption.

Corollary 1 (Stage-0 → 1 adoption condition)**.** *In the interior region, the delegated mode is used on a task iff*

$$\gamma + r_1 \;\leq\; \left(\sqrt{\frac{\rho s^2}{2\tau}} - (s\mu - X)\sqrt{\frac{\tau}{2\rho s^2}}\right)^2, \tag{16}$$

and under full delegation ($X \geq s\mu$) *iff* $\gamma + r_1 \leq X - s\mu + \rho s^2/(2\tau)$*. Reducing gating friction* γ *is the unique instrument that operates at every task simultaneously: capability growth raises* X *task by task, but friction reduction shifts the whole menu.*

(The right side of (16) equals $-(s\mu - X) + \rho s^2/(2\tau) + \tau(s\mu - X)^2/(2\rho s^2)$; the perfect-square form, verified in Appendix B, shows it is nonnegative.)

4.6 Menu nesting and its limits

Assumption 2 (Task separability)**.** Activation decisions are separable across tasks: no cross-task capacity constraint binds at either stage being compared.

Proposition 2 (Menu nesting)**.** *Let $\mathcal{M}_\ell \subseteq \mathcal{M}_{\ell+1}$ with the manual mode in every menu, and let $V^\ell = \max_{m\in\mathcal{M}_\ell} V^m$ and $Z^\ell = \mathbf{1}\{V^\ell \geq 0\}$. Under Assumption 2, $T^{\ell+1} \leq T^\ell$ task by task, hence $Z^{\ell+1} \geq Z^\ell$ and the number of active tasks weakly rises. The newly activated tasks are exactly those with $\omega \in [T^{\ell+1}, T^\ell)$, an event of probability $F(T^\ell) - F(T^{\ell+1})$.*

Example 2 (Why Assumption 2 is needed)**.** Let attention capacity be 1. Two tasks each yield value 1 and require attention 0.5 manually: both are done, task count 2, value 2. A new stage adds a mode making task A worth 5 but consuming attention 1. The optimum is A alone: value $5 > 2$ but task count $1 < 2$. Menus expanded and welfare rose, yet the task *count* fell. With binding capacity, the correct prediction is that value-weighted output rises, not that counts rise.

4.7 From task activation to the demand for validated changes

Order tasks by opportunity value. The number of tasks per period with value at least v is $D(v) = J\,(1 - F(v))$, so the Q-th most valuable task has value

$$\upsilon(Q) = F^{-1}\!\left(1 - \frac{Q}{J}\right), \qquad \upsilon'(Q) = -\frac{1}{Jf\big(F^{-1}(1 - Q/J)\big)} < 0. \tag{17}$$

Lemma 2 (Demand bridge)**.** *If F is continuous and strictly increasing, the gross value of the Q best tasks satisfies*

$$\Upsilon(Q) \equiv \int_0^Q \upsilon(u)\,du \;=\; J\int_{\upsilon(Q)}^{\bar\omega} \omega\,dF(\omega), \tag{18}$$

where $\bar\omega$ is the supremum of the support. The organization-level demand curve for validated changes is therefore the task-activation model itself: producing Q validated changes means completing precisely the tasks above the activation threshold $\upsilon(Q)$.

The change-of-variables proof is Appendix C. Two tractable special cases are used below: *linear demand* $\upsilon(Q) = \bar v - \varsigma Q$ and *box demand* $\upsilon(Q) = \bar v\,\mathbf{1}\{Q \leq \bar D\}$ (value $\bar v$ per change up to backlog $\bar D$, zero after).

5 The verification technology

5.1 Layered acceptance

An incorrect candidate is accepted iff it survives self-checks *and* is not caught downstream. With coverage g and review effort e,

$$q^{\mathrm{FA}}(x, g, e) = p^{\mathrm{sv}}(x)\left[(1 - g) + g\left(1 - R^G d(e)\right)\right] = p^{\mathrm{sv}}(x)\left[1 - g\,R^G d(e)\right]. \tag{19}$$

A correct candidate is accepted with probability $1 - g\,q^{\mathrm{FR}}(e)$.

Lemma 3 (Layered beats blended)**.** *Fix an external check that passes incorrect work with probability $p^{\mathrm{ext}} < 1$ and a self-check that passes it with probability $p^{\mathrm{sv}} < 1$. The blended (either/or) architecture*

yields false-pass probability $(1-g)p^{\mathrm{sv}} + g\,p^{\mathrm{ext}}$*; the layered architecture yields* $p^{\mathrm{sv}}\big[(1-g)+g\,p^{\mathrm{ext}}\big]$*. The difference is*

$$\underbrace{(1-g)p^{\mathrm{sv}} + g\,p^{\mathrm{ext}}}_{\textit{blended}} - \underbrace{p^{\mathrm{sv}}\big[(1-g)+g\,p^{\mathrm{ext}}\big]}_{\textit{layered}} = g\,p^{\mathrm{ext}}\,(1-p^{\mathrm{sv}}) \;\geq 0, \tag{20}$$

strict whenever $g \in (0,1)$*,* $p^{\mathrm{ext}} > 0$*, and* $p^{\mathrm{sv}} < 1$*. Serial checks multiply; substitutable checks average.*

Proposition 3 (Correlated-error floor)**.** *For every* x*,* $q^{\mathrm{FA}}(x,g,e) \geq \underline{p}(A)\big[1 - g\,R^G d(e)\big]$*, with the bound attained as* $x \to \infty$*. With no independent verification (*$g = 0$*),* $\inf_x q^{\mathrm{FA}} = \underline{p}(A) > 0$*: self-verification alone can never drive false acceptance below the correlated-error floor. Independent verification is necessary, not optional, for* $q^{\mathrm{FA}} < \underline{p}(A)$*.*

Proof. From (3), $p^{\mathrm{sv}}(x)$ is strictly decreasing with $\lim_{x\to\infty} p^{\mathrm{sv}}(x) = \underline{p}(A)$; substitute into (19). □

5.2 Guardrail reliability

Guardrails are only useful if they catch the failures that actually occur. Let h be the per-period rate at which the organization audits and updates its guardrail suite and δ_G the decay rate as the codebase and agent behavior drift. Appendix I derives the steady state

$$R^{G,*} = \frac{h}{h+\delta_G}, \tag{21}$$

which enters (19) directly: unaudited guardrails ($h \to 0$) drive $R^G \to 0$ and make coverage g ineffective against defects even at high review effort.

6 The organization's problem and the implementable scale

6.1 Review standards and flows

The organization sets a *review standard* $\hat{e}$: every externally reviewed candidate receives $\hat{e}$ hours. Candidates surviving self-checks arrive at rate

$$C^s = \zeta n\, S(x), \qquad S(x) \equiv \pi_c + (1-\pi_c)\,p^{\mathrm{sv}}(x), \tag{22}$$

since correct candidates (π_c) always survive and incorrect ones ($1-\pi_c$) survive with probability $p^{\mathrm{sv}}(x)$. Review capacity requires $g\,C^s\,\hat{e} \leq H^V$, i.e.

$$n \;\leq\; \bar{n}^V(x,g) \equiv \frac{H^V}{g\,\zeta\,S(x)\,\hat{e}}. \tag{23}$$

Isolation slots cap concurrent agents at

$$n \;\leq\; \bar{n}^I \equiv \min\left\{\frac{K^F}{m_F},\; \frac{K^R}{m_R}\right\}, \tag{24}$$

where m_F, m_R are per-agent filesystem and runtime slot requirements. Worktrees raise K^F only; ports, containers, databases, queues, and preview environments determine K^R, which is the binding branch whenever $K^R/m_R < K^F/m_F$. (Appendix H microfounds (24) with soft conflicts and shows

effective isolation aggregates *harmonically*, $1/I = m_F/K^F + m_R/K^R$, so effective capacity is below the weaker of the two.)

Accepted flows are

$$Q = \zeta n\, \pi_c \left(1 - g\, q^{\mathrm{FR}}(\hat{e})\right) \;=\; \mu_Q\, n, \qquad \mu_Q \equiv \zeta \pi_c \big(1 - g\, q^{\mathrm{FR}}(\hat{e})\big), \tag{25}$$

$$W = \zeta n\, (1 - \pi_c)\, p^{\mathrm{sv}}(x) \left(1 - g\, R^G d(\hat{e})\right) \;=\; \mu_W\, n, \quad \mu_W \equiv \zeta (1 - \pi_c)\, p^{\mathrm{sv}}(x) \big(1 - g\, R^G d(\hat{e})\big), \tag{26}$$

validated output and escaped defects respectively. Note $\mu_W = \zeta(1 - \pi_c)\, q^{\mathrm{FA}}$: escaped defects per agent are the false-acceptance probability applied to the incorrect flow.

6.2 The objective

Per-period profit is

$$\Pi(n, x, g) = \Upsilon(Q) - L\, W - (c_A + c_x x)\, \zeta n - w\, H^V - \sum_{m \in \{G,C,F,R\}} r_m K^m, \tag{27}$$

subject to (23), (24), $n \geq 0$, $x \geq 0$, $g \in [0, 1]$. Tokens ($\propto (1 + x)\zeta n$) are *usage*; Q is *validated output*; $\Upsilon(Q) - LW$ is *return*. The distinction between usage and return is an identity of the model, not an added proposition.

6.3 The implementable scale

Define the *per-agent margin* under box demand,

$$\pi_1 \equiv \bar{v}\, \mu_Q - L\, \mu_W - (c_A + c_x x)\, \zeta, \tag{28}$$

and the demand ceiling $\bar{n}^D \equiv \bar{D}/\mu_Q$ (the agent count at which validated output exhausts the backlog).

Assumption 3 (Regularity)**.** $\Pi(\cdot, x, g)$ is strictly quasiconcave in n on $[0, \infty)$ with peak $n^u \in (0, \infty]$. (Sufficient: v strictly decreasing, or box demand with $\pi_1 > 0$, in which case $n^u = \infty$.)

Theorem 1 (Implementable agent scale)**.** *Under Assumption 3 and the standards regime, the profit-maximizing feasible agent count is*

$$n^*(x, g) = \min\big\{n^u(x, g),\ \bar{n}^V(x, g),\ \bar{n}^I,\ \bar{n}^D(g)\big\}, \tag{29}$$

evaluated at the jointly optimal (x^*, g^*)*. If the per-agent margin is negative,* $n^* = 0$*.*

Proof. The feasible set is $[0, \bar{n}]$ with $\bar{n} = \min\{\bar{n}^V, \bar{n}^I, \bar{n}^D\}$ (each ceiling is a bound on n; $\bar{n}^D$ binds because Υ is flat beyond $\bar{D}$ under box demand, making additional agents pure cost). Strict quasiconcavity implies Π is strictly increasing on $[0, n^u)$; hence if $\bar{n} \leq n^u$ the maximum on $[0, \bar{n}]$ is at $\bar{n}$, and otherwise at n^u. Full steps in Appendix D. □

Three warnings accompany (29). First, the arms are *functions, not constants*: $\bar{n}^V$ rises with self-verification x and falls with coverage g, so the minimum must be evaluated at the optimal verification policy. Second, the formula requires genuinely *hard* constraints; if review dilution is smooth (Section 8) the verification ceiling becomes a smooth penalty and (29) is the Leontief limit. Third, under smooth demand one arm is redundant:

Corollary 2 (Redundant demand arm under smooth demand)**.** *With strictly decreasing υ and interior n^u, the first-order condition implies $\upsilon(Q(n^u)) = \big[L\mu_W + (c_A + c_x x)\zeta\big]/\mu_Q > 0$, so $Q(n^u) < Q(\bar{n}^D)$ and $n^u < \bar{n}^D$: the demand ceiling never binds separately. A distinct demand arm exists only under box demand.* (Proof: Appendix D.)

Proposition 4 (Interior scale under linear demand)**.** *With $\upsilon(Q) = \bar{v} - \varsigma Q$ and the standards regime,*

$$\Pi(n) = \bar{v}\mu_Q n - \frac{\varsigma}{2}\mu_Q^2 n^2 - L\mu_W n - (c_A + c_x x)\zeta n - wH^V - \textstyle\sum_m r_m K^m, \tag{30}$$

whose first-order condition $\bar{v}\mu_Q - \varsigma\mu_Q^2 n - L\mu_W - (c_A + c_x x)\zeta = 0$ yields

$$n^u = \frac{\bar{v}\,\mu_Q - L\,\mu_W - (c_A + c_x x)\,\zeta}{\varsigma\,\mu_Q^2}, \tag{31}$$

with second-order condition $-\varsigma\mu_Q^2 < 0$. (Steps: Appendix D.)

Remark 1 (Induced demand). If autonomous agents discover ϑ additional useful tasks per agent per period ($\bar{D}(n) = \bar{D}_0 + \vartheta n$ with $\vartheta < \mu_Q$), the demand ceiling solves $\mu_Q n = \bar{D}_0 + \vartheta n$:

$$\bar{n}^D = \frac{\bar{D}_0}{\mu_Q - \vartheta}. \tag{32}$$

Task discovery relaxes the demand ceiling but cannot remove it while $\vartheta < \mu_Q$; proposals in excess of useful demand are busywork, which belongs on the review-load side, not the value side.

6.4 Tokens are not enough

Proposition 5 (Bottleneck investment)**.** *Let $n^* = \bar{n}_k$ for exactly one binding arm $k \in \{V, I, D\}$, strictly below the other arms and n^u, with per-agent marginal profit $\pi_1 > 0$ at n^*. Then:*

(i) capital that relaxes only a slack arm has marginal value equal to $-r_m < 0$ (pure waste at the margin);

(ii) capital that relaxes the binding arm has marginal value $\pi_1\,\partial\bar{n}_k/\partial K^m - r_m$, which is positive when the relaxation rate is high enough;

(iii) raising raw generation ζ ("more tokens") when $k = V$ leaves validated output and profit unchanged *— the ceiling (23) shrinks n^* in exact inverse proportion — so any positive price paid for the increase is a strict loss.*

Remark 2 (Capability can tighten the verification ceiling). $\partial\bar{n}^V/\partial\pi_c = -H^V(1 - p^{\mathrm{sv}})/\big(g\zeta\hat{e}\,S^2\big) < 0$: higher raw correctness sends *more* survivors to review (correct candidates always survive self-checks), so when verification binds, a better model can *reduce* the supportable agent count even as value per candidate rises. Model upgrades and review-capacity investments are complements, not substitutes.

7 Optimal verification policy

7.1 Self-verification intensity

Proposition 6 (Optimal self-verification)**.** *In the standards regime the optimal self-verification intensity is*

$$x^* = \max\left\{0,\ \ln\left(\frac{L\,(1 - \pi_c)\,(p_0 - \underline{p})\,\big(1 - g\,R^G d(\hat{e})\big)}{c_x}\right)\right\}, \tag{33}$$

with strictly negative second-order condition. Comparative statics: x^* is increasing in the incident loss L and the error rate $1-\pi_c$, and decreasing in the harness cost c_x and in independent-verification coverage g ($\partial x^*/\partial g = -R^G d/(1-gR^G d) < 0$). Because $p^{\mathrm{SV}\prime}(x) \to 0$, optimal self-verification is finite even when free.

Derivation. The x-relevant terms of (27) are $-L\,\zeta n(1-\pi_c)p^{\mathrm{SV}}(x)\big(1-gR^G d(\hat{e})\big) - c_x x\,\zeta n$. Differentiating and using $p^{\mathrm{SV}\prime}(x) = -(p_0-\underline{p})e^{-x}$:

$$L\,(1-\pi_c)\,(p_0-\underline{p})\,e^{-x}\,\big(1-gR^G d(\hat{e})\big) = c_x, \tag{34}$$

whose solution is (33). Both sides scale with ζn, so x^* is independent of scale: self-verification is a per-candidate policy. Full steps in Appendix E. □

Equation (33) explains the stage pattern endogenously: at Stage 1 the harness is absent (c_x large), the log argument is below one, and $x^* = 0$ — the rational stage-1 team does *no* self-verification; as guardrail capital accumulates, c_x falls and x^* rises.

7.2 Coverage, substitutes, and complements

Holding $\hat{e}$ fixed, the cross-partial of the false-acceptance probability is

$$\frac{\partial^2 q^{\mathrm{FA}}}{\partial x\,\partial g} = -\,p^{\mathrm{SV}\prime}(x)\,R^G d(\hat{e}) = (p_0-\underline{p})e^{-x}R^G d(\hat{e}) > 0\,: \tag{35}$$

self- and independent verification are *substitutes in risk reduction* (each dulls the other's marginal effect). But through capacity they are *complements*: higher x shrinks the survivor flow (22), which relaxes the review constraint (23) and makes any coverage level cheaper to sustain:

Proposition 7 (Self-verification raises the verification ceiling)**.** $\dfrac{\partial \bar{n}^V}{\partial x} = \dfrac{H^V\,(1-\pi_c)(p_0-\underline{p})e^{-x}}{g\,\zeta\,\hat{e}\,S(x)^2} > 0$, *and* $\partial \bar{n}^V/\partial g < 0$, $\partial \bar{n}^V/\partial H^V > 0$.

This is the parallel-orchestration theorem: pre-review filtering is what converts one supervised agent into ten, because it acts on the constraint, not on the objective.

The coverage first-order condition balances the value of catching defects against false rejection of correct work:

$$\frac{\partial \Pi}{\partial g} = \zeta n\Big[L\,(1-\pi_c)\,p^{\mathrm{SV}}(x)\,R^G d(\hat{e}) \;-\; \upsilon(Q)\,\pi_c\,q^{\mathrm{FR}}(\hat{e})\Big] \;-\; (\text{shadow cost of the capacity constraint}), \tag{36}$$

with corner $g^* = 1$ when incident losses dominate and interior g^* when review capacity or false rejection bites. Exception-based review at high stages is exactly a fall in optimal coverage as guardrails ($R^G d$) take over the routine layer — the knowledge-hierarchy structure of [11].

8 Soft utilization: risk rises with scale

Replace the standards regime with a fixed reviewer pool spread over whatever arrives: effort per reviewed candidate becomes the *utilization* outcome

$$e(n) = \frac{H^V}{g\,\zeta\,n\,S(x)}, \qquad \frac{\partial e}{\partial n} = -\frac{e}{n} < 0, \qquad \frac{\partial e}{\partial x} = e\,\frac{(1-\pi_c)(p_0-\underline{p})e^{-x}}{S(x)} > 0. \tag{37}$$

Lemma 4 (Dilution). *For any increasing concave d with $d(0) = 0$, $\psi(e) \equiv d(e) - e\,d'(e) \geq 0$, with $\psi(0) = 0$ and $\psi'(e) = -e\,d''(e) \geq 0$.*

Proposition 8 (The marginal agent dilutes review). *Under (37),*

$$\frac{dQ}{dn} = \zeta\pi_c\Big[\big(1 - g\,q^{\mathrm{FR}}(e)\big) + g\,q^{\mathrm{FR}\prime}(e)\,e\Big], \tag{38}$$

$$\frac{dW}{dn} = \zeta(1-\pi_c)\,p^{\mathrm{SV}}(x)\Big[\big(1 - g\,R^G d(e)\big) + g\,R^G d'(e)\,e\Big]. \tag{39}$$

The bracketed second terms are the utilization externality*: since $q^{\mathrm{FR}\prime} < 0$, the marginal agent lowers everyone's validated yield; since $d' > 0$, it raises everyone's escaped-defect rate. Consequently q^{FA} is* increasing in n*: risk depends on review intensity per candidate, not on total review capacity. Coverage still weakly reduces false acceptance under dilution, because $\frac{\partial}{\partial g}\big[1 - gR^G d(e(g))\big] = -R^G\psi(e) \leq 0$ by Lemma 4.* (Derivations: Appendix F.)

The hard ceiling (23) is the limit of this soft technology as d approaches a step function at $\hat{e}$; the minimum formula (29) is exact in that limit and an approximation otherwise.

9 Stages as binding constraints

Definition 5 (Stage). Given a configuration $\sigma = (\gamma, x, g, \hat{e}, K^G, K^C, K^F, K^R, H^V)$, the *stage* is the identity of the binding term in (29), with Stage 0 the case $n^* = 0$ (friction γ or a negative margin blocks participation), and higher stages ordered by the equilibrium scale n^*.

Table 1 evaluates the model at four configurations chosen to mirror the practitioner ladder. Parameters common to all rows: $H^V = 40$ review hours/week, $\zeta = 5$ candidates/agent/week, $\bar{v} = 100$, $c_A = 2$, $p_0 = 0.9$, $\underline{p} = 0.02$. Self-verification is set at its optimum (33) row by row; $p^{\mathrm{SV}} = p^{\mathrm{SV}}(x^*)$. All values below are computed from equations (3), (33), (22)–(28), and (29).

The ladder magnitudes $1 \to 10 \to 100 \to 1000$ emerge without being assumed, and the binding constraint migrates: line-by-line review caps Stage 1 at one agent; at Stage 2 self-verification lifts the review ceiling to eleven but *runtime* isolation (ports, databases, services — not worktrees, which are ample at 20) caps scale at eight; at Stage 3 containers relax isolation and exception-based review (low g, high R^G) supports about one hundred agents with verification binding again; at Stage 4 the constraint is *useful demand* — the organization can verify and isolate more agents than it has valuable work for, which is why AI-native operation is characterized by intent-setting and task discovery (ϑ in Remark 1) rather than by further verification investment. The stage-by-stage rise of x^* (0, 1.26, 2.42, 4.47) is driven by the falling harness cost $c_x(K^G)$, not by assumption.

10 Stage transitions

Let Π_ℓ denote maximized per-period profit under configuration σ_ℓ (rentals $r_m K^m$ included as flows), and let $\Phi_{\ell+1}$ be the *one-time installation cost* of the capital increment that defines the next configuration. Installation and rental are distinct: Φ prices the transition once; $r_m K^m$ prices holding the stock each period. No expenditure is counted twice.

Proposition 9 (Transition rule). *With discount factor $\beta \in (0,1)$ and profit gains starting the period after installation, moving from ℓ to $\ell+1$ is worthwhile iff*

$$\sum_{r=1}^{\infty} \beta^r \left(\Pi_{\ell+1} - \Pi_\ell\right) = \frac{\beta}{1-\beta}\left(\Pi_{\ell+1} - \Pi_\ell\right) \;\geq\; \Phi_{\ell+1}, \qquad \textit{i.e.} \qquad \Pi_{\ell+1} - \Pi_\ell \;\geq\; \frac{1-\beta}{\beta}\,\Phi_{\ell+1}. \tag{40}$$

Table 1: The stage ladder as bottleneck migration (all rows at optimal x^*)

	Stage 1	Stage 2	Stage 3	Stage 4
π_c (correct rate)	0.60	0.60	0.80	0.95
g (coverage)	1	1	0.10	0.005
$\hat{e}$ (hours/review)	8	1	1	1
$R^G d(\hat{e})$	0.40	0.25	0.45	0.475
$q^{\mathrm{FR}}(\hat{e})$	0.05	0.10	0.10	0.10
c_x (harness cost)	60	15	3	1
L (incident loss)	200	200	200	2000
K^F/m_F, K^R/m_R	3, 2	20, 8	500, 120	5000, 2000
$\bar{D}$ (backlog/week)	200	200	600	5000
x^* (33)	0	1.26	2.42	4.47
$p^{\mathrm{sv}}(x^*)$	0.90	0.27	0.10	0.03
$S(x^*)$	0.96	0.71	0.82	0.95
$\bar{n}^V$ (23)	1.0	11.3	97.6	1682
$\bar{n}^I$ (24)	2	8	120	2000
$\bar{n}^D$	70	74	152	1053
per-agent margin π_1 (28)	59	85	331	427
n^* (29)	≈ 1	8	≈ 98	≈ 1053
binding constraint	verification	**runtime isolation**	verification	**demand**

If capability A_t grows over time and $\Pi_{\ell+1} - \Pi_\ell$ is increasing in A, adoption follows a threshold date $t^*_{\ell+1} = \min\{t : \Pi_{\ell+1}(A_t) - \Pi_\ell(A_t) \geq \frac{1-\beta}{\beta}\Phi_{\ell+1}\}$, and stages are adopted in order. (Proof: Appendix G.)

Remark 3 (Complementary capital, not tokens). By Proposition 5, $\Pi_{\ell+1} - \Pi_\ell$ is generated only by investments that relax the *binding* arm of (29) at stage ℓ: access friction γ at Stage 0, review capacity and harness cost at Stage 1, runtime isolation at Stage 2, guardrail reliability and coverage policy at Stage 3, task discovery at Stage 4. Buying capability or tokens while the binding arm is unchanged leaves $\Pi_{\ell+1} = \Pi_\ell$ and fails (40) at any $\Phi > 0$.

Remark 4 (Option value). Under uncertainty about A_t or demand, (40) is the zero-volatility limit; irreversible Φ adds an option value of waiting that delays adoption beyond the NPV threshold [12].

11 Empirical content

11.1 Mapping model objects to telemetry

Model object	Observable proxy
n	concurrent agent sessions; active parallel branches per developer-day
x	agent-initiated test/build/lint runs per candidate before the pull request
g	share of agent PRs with required human review or protected-branch gates
$\hat{e}$, e	reviewer-minutes per reviewed PR; review rounds
H^V	total reviewer-hours per period

K^G, R^G	test coverage, CI gates, policy-as-code counts; fraction of injected defects caught in guardrail audits; guardrail-update commit rate h
K^C	size and churn of shared rules files, prompt libraries, evals, runbooks
K^F, K^R	worktrees per repository; containers, namespaces, preview environments
C, Q, W	candidate PRs opened; merged PRs; post-merge reverts, hotfixes, incident-linked PRs
q^{FA}	revert/incident rate among merged agent PRs
q^{FR}	rejected-then-merged-unchanged rate; reopened candidates
$\upsilon(\cdot)$, $\bar{D}$, ϑ	backlog depth and priority distribution; share of merged work never linked to a pre-existing issue
γ	time-to-approval for AI tooling; procurement and security-review lead times
usage vs. return	tokens and API spend vs. $\Upsilon(Q) - LW$

11.2 Predictions

(P1) *Role substitution.* λ^* rises with capability (11): execution decisions shift to agents while planning and verification decisions remain human, as observed in usage telemetry [3].

(P2) *Delegation heterogeneity.* More risk-averse developers delegate more; developers with longer tenure on a codebase (higher τ) delegate less (Proposition 1).

(P3) *Self-verification scales parallelism.* Teams with pre-review CI harnesses (low c_x, high x^*) support more concurrent sessions per reviewer (Proposition 7); the elasticity works through the survivor share $S(x)$.

(P4) *Bottleneck migration.* The diagnostic that predicts the next investment differs by stage: review-queue saturation (verification-bound), merge conflicts and port/database collisions (isolation-bound), idle agent capacity and falling marginal task value (demand-bound). Which diagnostic binds is observable from the same telemetry as n^*.

(P5) *Task variety.* Stage advancement widens the set of task types performed — but only where no cross-task capacity binds (Proposition 2 and Example 2); under binding review capacity, value-weighted output rises while counts may not.

(P6) *Usage is not return.* When a non-demand arm binds, additional tokens raise usage and strictly lower profit (Proposition 5(iii)); the empirical implication is to report validated output, review minutes per accepted change, revert rates, and incident costs, never token counts alone.

12 Conclusion

The stages of agentic automation are not a capability dial; they are a sequence of binding constraints. One model — delegation at the task level, layered verification with a correlated-error floor, and an organization-level scale choice against verification, isolation, and demand ceilings — reproduces the practitioner ladder as the migration path of the binding arm of a single minimum formula, and prices every transition as an investment in the bottleneck it breaks. The model's sharpest lessons are structural: self-verification operates on the *constraint* (it lifts the review ceiling) while independent verification operates on the *objective* (it is the only instrument that beats the correlated-error

floor); capability can tighten the verification ceiling even as it raises value; and at the top of the ladder the scarce resource is neither attention nor compute but useful demand. Extensions in the appendix cover guardrail maintenance dynamics, codification and the diffusion gap behind the "10x developer," regulatory burdens, and a workflow-level readiness index for organizations whose teams sit at different stages. Business-model frictions (hourly billing that resists automation gains) are left for future work.

A The CARA–Normal certainty equivalent

Let $Y \mid \mathcal{F} \sim N(m, v)$ and $u(y) = -e^{-\rho y}$, $\rho > 0$. The moment generating function of a Normal random variable is $\mathbb{E}[e^{qY}] = \exp(qm + \frac{1}{2}q^2 v)$. Set $q = -\rho$:

$$\mathbb{E}[e^{-\rho Y}] = \exp\left(-\rho m + \frac{\rho^2 v}{2}\right).$$

Hence expected utility is

$$\mathbb{E}[u(Y)] = -\mathbb{E}[e^{-\rho Y}] = -\exp\left(-\rho m + \frac{\rho^2 v}{2}\right).$$

The certainty equivalent solves $-e^{-\rho\, CE} = \mathbb{E}[u(Y)]$. Cancel the minus signs and take logarithms:

$$-\rho\, CE = -\rho m + \frac{\rho^2 v}{2}.$$

Divide by $-\rho$:

$$CE = m - \frac{\rho}{2} v. \qquad \blacksquare$$

B Proofs for Section 4

B.1 Proposition 1 (optimal delegation)

Start from (10):

$$V^{\text{del}}(\lambda) = \omega + (1-\lambda)s\mu + \lambda X - \gamma - r_1 - b - \frac{\rho}{2}\frac{(1-\lambda)^2 s^2}{\tau}.$$

Step 1 (first derivative). Differentiate term by term. $\frac{d}{d\lambda}[(1-\lambda)s\mu] = -s\mu$; $\frac{d}{d\lambda}[\lambda X] = X$; for the risk term, write $(1-\lambda)^2 = 1 - 2\lambda + \lambda^2$, so

$$\frac{d}{d\lambda}\left[-\frac{\rho s^2}{2\tau}(1-\lambda)^2\right] = -\frac{\rho s^2}{2\tau}\cdot 2(1-\lambda)\cdot(-1) = +\frac{\rho s^2}{\tau}(1-\lambda).$$

Collecting,

$$\frac{dV^{\text{del}}}{d\lambda} = -s\mu + X + \frac{\rho s^2}{\tau}(1-\lambda).$$

Step 2 (second derivative). $\frac{d^2 V^{\text{del}}}{d\lambda^2} = -\rho s^2/\tau < 0$: strictly concave, so the first-order condition characterizes the maximum, clipped to $[0, 1]$.

Step 3 (interior solution). Set the first derivative to zero:

$$\frac{\rho s^2}{\tau}(1-\lambda) = s\mu - X \quad\Longrightarrow\quad 1 - \lambda^* = \frac{\tau\,(s\mu - X)}{\rho s^2} \quad\Longrightarrow\quad \lambda^* = 1 - \frac{\tau\,(s\mu - X)}{\rho s^2}.$$

Step 4 (corners). $\lambda^* \geq 1$ iff $s\mu - X \leq 0$, i.e. $X \geq s\mu$ (full delegation); $\lambda^* \leq 0$ iff $\tau(s\mu - X) \geq \rho s^2$, i.e. $X \leq s\mu - \rho s^2/\tau$ (no delegation).
Step 5 (comparative statics, interior). With $y \equiv s\mu - X > 0$ in the interior region:

$$\frac{\partial \lambda^*}{\partial \rho} = \frac{\tau y}{\rho^2 s^2} > 0, \qquad \frac{\partial \lambda^*}{\partial \tau} = -\frac{y}{\rho s^2} < 0, \qquad \frac{\partial \lambda^*}{\partial A} = \frac{\tau}{\rho s^2}\, a\, z'(A) > 0,$$

the last using $\partial X/\partial A = a\, z'(A) > 0$ from (9). ■

B.2 Envelope value (13)

Substitute $1 - \lambda^* = \tau y/(\rho s^2) \equiv \Delta$ (interior case) into (10):

$$V^{\mathrm{del},*} = \omega + \Delta\, s\mu + (1-\Delta)X - \gamma - r_1 - b - \frac{\rho s^2}{2\tau}\Delta^2.$$

Step 1. $\Delta s\mu + (1-\Delta)X = X + \Delta(s\mu - X) = X + \Delta y$. **Step 2.** $\Delta y = \frac{\tau y}{\rho s^2} \cdot y = \frac{\tau y^2}{\rho s^2}$. **Step 3.** $\frac{\rho s^2}{2\tau}\Delta^2 = \frac{\rho s^2}{2\tau} \cdot \frac{\tau^2 y^2}{\rho^2 s^4} = \frac{\tau y^2}{2\rho s^2}$. **Step 4.** Combining, $V^{\mathrm{del},*} = \omega + X + \frac{\tau y^2}{\rho s^2} - \frac{\tau y^2}{2\rho s^2} - \gamma - r_1 - b = \omega + X + \frac{\tau(s\mu - X)^2}{2\rho s^2} - \gamma - r_1 - b$, which is (13). Setting this to zero gives the threshold (14). At $\lambda^* = 1$ the value is $\omega + X - \gamma - r_1 - b$ directly; at $\lambda^* = 0$ it is $V^{\mathrm{man}} - \gamma - r_1$, dominated by the manual mode. ■

B.3 Corollary 1 (perfect-square adoption condition)

The delegated mode is used iff $T^{\mathrm{del}} \leq T^{\mathrm{man}}$. Using (7) and (14), this is

$$b + \gamma + r_1 - X - \frac{\tau y^2}{2\rho s^2} \;\leq\; b - s\mu + \frac{\rho s^2}{2\tau} \quad\Longleftrightarrow\quad \gamma + r_1 \;\leq\; -y + \frac{\rho s^2}{2\tau} + \frac{\tau y^2}{2\rho s^2},$$

using $X - s\mu = -y$. Let $u = \sqrt{\rho s^2/(2\tau)}$ and $w = y\sqrt{\tau/(2\rho s^2)}$. Then $u^2 = \rho s^2/(2\tau)$, $w^2 = \tau y^2/(2\rho s^2)$, and $2uw = 2y\sqrt{1/4} = y$, so the right side equals $u^2 - 2uw + w^2 = (u-w)^2$, which is (16). ■

B.4 Proposition 2 (menu nesting)

Since $\mathcal{M}_\ell \subseteq \mathcal{M}_{\ell+1}$, $V^{\ell+1} = \max_{m \in \mathcal{M}_{\ell+1}} V^m \geq \max_{m \in \mathcal{M}_\ell} V^m = V^\ell$ task by task; under Assumption 2 this maximization is unconstrained across tasks, so $V^\ell \geq 0$ implies $V^{\ell+1} \geq 0$, giving $Z^{\ell+1} \geq Z^\ell$ and, summing, weakly more active tasks. Every mode's value is additive in ω with unit coefficient, so each V^m is strictly increasing in ω and each stage value V^ℓ crosses zero once, at the threshold $T^\ell = \min_{m \in \mathcal{M}_\ell} T^m$; menu expansion weakly lowers the minimum. The activation band and its probability follow from $T^{\ell+1} \leq T^\ell$ and the definition of the CDF. ■

C Proof of Lemma 2 (demand bridge)

By definition $\Upsilon(Q) = \int_0^Q F^{-1}(1 - u/J)\, du$. Substitute $u = J\,(1 - F(\omega))$, so that $\omega = F^{-1}(1 - u/J)$ and $du = -Jf(\omega)\, d\omega$. When $u = 0$, $\omega = \bar{\omega}$; when $u = Q$, $\omega = \upsilon(Q)$. Then

$$\Upsilon(Q) = \int_{\bar{\omega}}^{\upsilon(Q)} \omega\, \big(-Jf(\omega)\big)\, d\omega = J \int_{\upsilon(Q)}^{\bar{\omega}} \omega f(\omega)\, d\omega,$$

which is (18). (Check, uniform F on $[0,1]$, J arbitrary: left side $\int_0^Q (1-u/J)du = Q - Q^2/(2J)$; right side $J\int_{1-Q/J}^1 \omega\, d\omega = \frac{J}{2}\left[1-(1-Q/J)^2\right] = Q - Q^2/(2J)$.) ■

D Proofs for Section 6

D.1 Theorem 1

Step 1 (the feasible set). Constraint (23) bounds n by $\bar{n}^V$; constraint (24) bounds n by $\bar{n}^I$; under box demand, output beyond $\bar{D}$ has zero value while costs are strictly positive, so no maximizer exceeds $\bar{n}^D = \bar{D}/\mu_Q$. Hence the search can be restricted to $[0, \bar{n}]$, $\bar{n} \equiv \min\{\bar{n}^V, \bar{n}^I, \bar{n}^D\}$.
Step 2 (single-peakedness). By Assumption 3, Π is strictly quasiconcave with peak n^u: strictly increasing on $[0, n^u)$ and strictly decreasing on (n^u, ∞).
Step 3 (cases). If $n^u \le \bar{n}$, the unconstrained peak is feasible and uniquely optimal: $n^* = n^u$. If $n^u > \bar{n}$, Π is strictly increasing throughout $[0, \bar{n}]$, so the maximum is at the boundary: $n^* = \bar{n}$. In both cases $n^* = \min\{n^u, \bar{n}\} = \min\{n^u, \bar{n}^V, \bar{n}^I, \bar{n}^D\}$, which is (29). If the per-agent margin is negative at $n = 0^+$, Π is decreasing from the start and $n^* = 0$. ■

D.2 Corollary 2

At an interior peak with strictly decreasing υ, the first-order condition of (27) in n (standards regime, so μ_Q, μ_W are constants) reads

$$\upsilon\big(Q(n^u)\big)\,\mu_Q - L\,\mu_W - (c_A + c_x x)\,\zeta = 0 \quad\Longrightarrow\quad \upsilon\big(Q(n^u)\big) = \frac{L\,\mu_W + (c_A + c_x x)\,\zeta}{\mu_Q} > 0.$$

Since υ is strictly decreasing and $\upsilon(Q(\bar{n}^D)) = 0$ by definition of the demand ceiling, $Q(n^u) < Q(\bar{n}^D)$, hence $n^u < \bar{n}^D$: the demand arm is slack whenever the interior peak exists. ■

D.3 Proposition 4 (interior scale, linear demand)

With $\upsilon(Q) = \bar{\upsilon} - \varsigma Q$, $\Upsilon(Q) = \int_0^Q (\bar{\upsilon} - \varsigma u)du = \bar{\upsilon}Q - \frac{\varsigma}{2}Q^2$. Substitute $Q = \mu_Q n$:

$$\Pi(n) = \bar{\upsilon}\mu_Q n - \frac{\varsigma}{2}\mu_Q^2 n^2 - L\mu_W n - (c_A + c_x x)\zeta n - wH^V - \textstyle\sum_m r_m K^m.$$

Step 1. Differentiate: $\Pi'(n) = \bar{\upsilon}\mu_Q - \varsigma\mu_Q^2 n - L\mu_W - (c_A + c_x x)\zeta$. **Step 2.** Set to zero and solve:

$$\varsigma\mu_Q^2\, n = \bar{\upsilon}\mu_Q - L\mu_W - (c_A + c_x x)\zeta \quad\Longrightarrow\quad n^u = \frac{\bar{\upsilon}\mu_Q - L\mu_W - (c_A + c_x x)\zeta}{\varsigma\mu_Q^2}.$$

Step 3. $\Pi''(n) = -\varsigma\mu_Q^2 < 0$. ■

D.4 Proposition 5 (bottleneck investment)

At $n^* = \bar{n}_k$ strictly below all other arms and n^u, small changes in a slack arm leave the feasible boundary $\bar{n} = \bar{n}_k$ unchanged, so the only effect of capital K^m that relaxes a slack arm is its rental: $d\Pi^*/dK^m = -r_m < 0$, part (i). For capital relaxing the binding arm, $d\Pi^*/dK^m = \Pi'(\bar{n}_k)\cdot \partial\bar{n}_k/\partial K^m - r_m = \pi_1\,\partial\bar{n}_k/\partial K^m - r_m$, part (ii), where $\Pi'(\bar{n}_k) = \pi_1 > 0$ because the peak lies beyond the boundary. For part (iii): when $k = V$, $n^* = \bar{n}^V = H^V/(g\zeta S\hat{e})$, so the product $\zeta n^* = H^V/(gS\hat{e})$ is invariant to ζ. Every flow is a function of ζn only: $Q = \pi_c(1 - gq^{\mathrm{FR}})\,\zeta n^*$, $W = (1-\pi_c)p^{\mathrm{sv}}(1 - gR^G d)\,\zeta n^*$, and compute cost $(c_A + c_x x)\,\zeta n^*$ are all unchanged. Hence Π^* is unchanged by the increase in ζ, and if the increase is acquired at any positive price, profit strictly falls; more tokens at a binding verification ceiling raise usage per agent, not validated output. ■

E Proof for Section 7

E.1 Proposition 6

The x-dependent part of (27) in the standards regime is

$$\Pi_x = -L\,\zeta n\,(1-\pi_c)\,p^{\mathrm{SV}}(x)\,\big(1-gR^G d(\hat{e})\big) - c_x\,x\,\zeta n.$$

Step 1. From (3), $p^{\mathrm{SV}\prime}(x) = -(p_0-\underline{p})e^{-x}$. **Step 2.** Differentiate:

$$\frac{d\Pi_x}{dx} = L\,\zeta n\,(1-\pi_c)\,(p_0-\underline{p})e^{-x}\big(1-gR^G d(\hat{e})\big) - c_x\zeta n.$$

Step 3. Set to zero and divide by $\zeta n > 0$:

$$L\,(1-\pi_c)\,(p_0-\underline{p})\,e^{-x}\big(1-gR^G d(\hat{e})\big) = c_x \implies e^{-x} = \frac{c_x}{L(1-\pi_c)(p_0-\underline{p})\big(1-gR^G d(\hat{e})\big)},$$

and taking logarithms gives (33); the corner $x^* = 0$ applies when the log argument is at most one. **Step 4 (second order).** $\dfrac{d^2\Pi_x}{dx^2} = -L\zeta n(1-\pi_c)(p_0-\underline{p})e^{-x}\big(1-gR^G d(\hat{e})\big) < 0$. **Step 5 (statics).** Write $x^* = \ln L + \ln(1-\pi_c) + \ln(p_0-\underline{p}) + \ln(1-gR^G d) - \ln c_x$. Then $\partial x^*/\partial L = 1/L > 0$, $\partial x^*/\partial c_x = -1/c_x < 0$, and

$$\frac{\partial x^*}{\partial g} = \frac{-R^G d(\hat{e})}{1-gR^G d(\hat{e})} < 0. \qquad \blacksquare$$

E.2 Proposition 7

From (23), $\bar{n}^V = H^V/(g\zeta\hat{e}\,S(x))$ with $S(x) = \pi_c + (1-\pi_c)p^{\mathrm{SV}}(x)$. **Step 1.** $S'(x) = (1-\pi_c)p^{\mathrm{SV}\prime}(x) = -(1-\pi_c)(p_0-\underline{p})e^{-x} < 0$. **Step 2.** By the quotient rule,

$$\frac{\partial \bar{n}^V}{\partial x} = -\frac{H^V}{g\zeta\hat{e}}\cdot\frac{S'(x)}{S(x)^2} = \frac{H^V(1-\pi_c)(p_0-\underline{p})e^{-x}}{g\zeta\hat{e}\,S(x)^2} > 0.$$

Step 3. $\partial\bar{n}^V/\partial g = -H^V/(g^2\zeta\hat{e}S) < 0$ and $\partial\bar{n}^V/\partial H^V = 1/(g\zeta\hat{e}S) > 0$ directly. $\blacksquare$

F Proofs for Section 8

F.1 Lemma 4

Let $\psi(e) = d(e) - e\,d'(e)$. Then $\psi(0) = d(0) - 0 = 0$ and $\psi'(e) = d'(e) - d'(e) - e\,d''(e) = -e\,d''(e) \geq 0$ by concavity, so $\psi(e) \geq 0$ for $e \geq 0$. $\blacksquare$

F.2 Proposition 8

With $e(n) = H^V/(g\zeta n S)$ we have $e'(n) = -H^V/(g\zeta n^2 S) = -e/n$. **Validated flow.** $Q(n) = \zeta n\pi_c\big(1 - g\,q^{\mathrm{FR}}(e(n))\big)$; by the product and chain rules,

$$\frac{dQ}{dn} = \zeta\pi_c\big(1-gq^{\mathrm{FR}}(e)\big) + \zeta n\pi_c\cdot\big(-g\,q^{\mathrm{FR}\prime}(e)\big)\cdot e'(n) = \zeta\pi_c\Big[\big(1-gq^{\mathrm{FR}}(e)\big) + g\,q^{\mathrm{FR}\prime}(e)\,e\Big],$$

using $n\,e'(n) = -e$; since $q^{\mathrm{FR}\prime} < 0$, the dilution term is negative. **Escaped defects.** $W(n) = \zeta n(1-\pi_c)p^{\mathrm{sv}}\big(1 - gR^G d(e(n))\big)$; identically,

$$\frac{dW}{dn} = \zeta(1-\pi_c)p^{\mathrm{sv}}\Big[\big(1 - gR^G d(e)\big) + g\,R^G d'(e)\,e\Big],$$

and since $d' > 0$ the dilution term is positive: the marginal agent raises the escape rate of *all* candidates. Because $e(n)$ is decreasing and d increasing, $q^{\mathrm{FA}}(x, g, e(n)) = p^{\mathrm{sv}}(x)\big[1 - gR^G d(e(n))\big]$ is increasing in n. **Coverage under dilution.** With $e(g) = H^V/(g\zeta nS)$, $\partial e/\partial g = -e/g$, so

$$\frac{\partial}{\partial g}\big[1 - gR^G d(e(g))\big] = -R^G d(e) - gR^G d'(e)\left(-\frac{e}{g}\right) = -R^G\big[d(e) - e\,d'(e)\big] = -R^G\psi(e) \le 0$$

by Lemma 4. ■

G Proof of Proposition 9

Staying at ℓ forever from date $t+1$ has present value $\sum_{r\ge1}\beta^r\Pi_\ell = \frac{\beta}{1-\beta}\Pi_\ell$. Paying $\Phi_{\ell+1}$ at t and operating at $\ell+1$ from $t+1$ has present value $-\Phi_{\ell+1} + \frac{\beta}{1-\beta}\Pi_{\ell+1}$. Switching is worthwhile iff

$$-\Phi_{\ell+1} + \frac{\beta}{1-\beta}\Pi_{\ell+1} \;\ge\; \frac{\beta}{1-\beta}\Pi_\ell \quad\Longleftrightarrow\quad \frac{\beta}{1-\beta}\big(\Pi_{\ell+1} - \Pi_\ell\big) \;\ge\; \Phi_{\ell+1},$$

which is (40); the geometric sum uses $\sum_{r=1}^{\infty}\beta^r = \beta/(1-\beta)$ for $\beta \in (0,1)$. If $\Delta\Pi(A_t) \equiv \Pi_{\ell+1}(A_t) - \Pi_\ell(A_t)$ is increasing in A_t and A_t is increasing in t, the set of dates satisfying (40) is an upper interval, so adoption occurs at its first element $t^*_{\ell+1}$ and thresholds are crossed in order of stage. ■

H Soft isolation: microfoundation

Suppose each running agent, over a period, collides with another agent's *filesystem* state as a Poisson event with hazard n/I^F, $I^F = K^F/m_F$, and with shared *runtime* resources (ports, containers, databases, queues) with hazard n/I^R, $I^R = K^R/m_R$, the two independent conditional on n. The probability a candidate completes without conflict is

$$\mathbb{P}(\text{no conflict}) = e^{-n/I^F}\,e^{-n/I^R} = \exp\left[-n\left(\frac{1}{I^F} + \frac{1}{I^R}\right)\right] \equiv e^{-n/I^{\mathrm{eff}}}, \qquad \frac{1}{I^{\mathrm{eff}}} = \frac{m_F}{K^F} + \frac{m_R}{K^R}. \quad (41)$$

Effective isolation aggregates *harmonically*: $I^{\mathrm{eff}} < \min\{I^F, I^R\}$, and when worktrees make I^F large, $I^{\mathrm{eff}} \approx I^R$ — runtime isolation binds. Completed candidates are $C(n) = \zeta n\,e^{-n/I^{\mathrm{eff}}}$. With value v_c per completed candidate and compute c_A per launched candidate, the scale choice solves $\max_n\; v_c\zeta n e^{-n/I^{\mathrm{eff}}} - c_A\zeta n$, with first-order condition

$$v_c\,e^{-n/I^{\mathrm{eff}}}\left(1 - \frac{n}{I^{\mathrm{eff}}}\right) = c_A.$$

Uniqueness. The left side has derivative $-\frac{v_c}{I^{\mathrm{eff}}}e^{-n/I^{\mathrm{eff}}}\big(2 - n/I^{\mathrm{eff}}\big) < 0$ for $n < 2I^{\mathrm{eff}}$, so it falls monotonically from v_c at $n=0$ to 0 at $n = I^{\mathrm{eff}}$, and the solution is unique in $(0, I^{\mathrm{eff}})$ for $c_A/v_c \in (0,1)$. **Limit.** As $c_A/v_c \to 0$, $n \to I^{\mathrm{eff}}$: with cheap compute, *optimal agent count equals effective isolation capacity*. The hard cap (24) is the limiting case in which conflict probability jumps from zero to one at the slot capacity.

I Dynamic extensions

I.1 Guardrail reliability

Let R_t^G evolve as $R_{t+1}^G = R_t^G + h\,(1 - R_t^G) - \delta_G R_t^G$: auditing repairs a share h of the gap to full reliability; drift erodes a share δ_G of current reliability. The steady state solves

$$R^G = R^G + h(1 - R^G) - \delta_G R^G \implies 0 = h - (h + \delta_G)R^G \implies R^{G,*} = \frac{h}{h + \delta_G},$$

which is (21). Because R^G multiplies $d(e)$ in (19), guardrail maintenance is a first-order input to safe scale: an organization that builds gates but does not audit them ($h \to 0$) operates with $R^G \to 0$ and obtains no defect protection from coverage.

I.2 Codification and the diffusion gap

Codification capital accumulates as $K_{t+1}^C = K_t^C + \phi_d\, d_t - \delta_K K_t^C$, where d_t is documentation and standardization effort. Candidate productivity $\zeta(A, K^C)$ and correctness $\pi_c(A, K^C)$ rise with K^C. The "10x developer" phenomenon is the gap

$$\Delta^{\text{diff}} = \zeta\big(A, K^{C,\text{private}}\big) - \zeta\big(A, K^{C,\text{shared}}\big)\,:$$

one person owns private prompts, context, and review habits that the team has not inherited. Team-level stage advancement requires raising $K^{C,\text{shared}}$, which simultaneously raises every agent's ζ and π_c and lowers effective review cost.

I.3 Regulatory and legacy burden

For a workflow with regulatory burden $\varrho \geq 0$: review cost $\kappa_1(\varrho) = \kappa_1 e^{\xi_\kappa \varrho}$, incident loss $L(\varrho) = L e^{\xi_L \varrho}$, and installation cost $\Phi_{\ell+1}(\varrho) = \Phi_{\ell+1} + \xi_\Phi \varrho$, with $\xi_\kappa, \xi_L, \xi_\Phi \geq 0$. All three channels lower X, raise x^* and optimal coverage, shrink n^*, and delay $t_{\ell+1}^*$: the same organization is rationally at different stages in regulated and unregulated workflows.

I.4 The ragged frontier

Grade each workflow w on three axes against stage-specific standards: autonomy $\ell_w^{\text{aut}} = \max\{\ell : \lambda_w \geq \bar{\lambda}_\ell\}$, parallelism $\ell_w^{\text{par}} = \max\{\ell : n_w^* \geq \bar{n}_\ell\}$, and codification $\ell_w^{\text{cod}} = \max\{\ell : K_w^C \geq \bar{K}_\ell\}$. The workflow's stage readiness is the minimum of the three *ordinal grades*,

$$\ell_w = \min\big\{\ell_w^{\text{aut}},\ \ell_w^{\text{par}},\ \ell_w^{\text{cod}}\big\},$$

(the minimum is taken over commensurable integer grades, never over raw quantities in different units), and the organization is described by the distribution $M_\ell = \frac{1}{|\mathcal{W}|}\sum_w \mathbf{1}\{\ell_w \geq \ell\}$ rather than by a scalar maturity level.

J Notation

Symbol	Meaning
$i,\ j,\ t$	team, task, period

$\omega,\ F,\ J,\ b$	task opportunity value, its distribution, task mass, activation cost
$s,\ a$	task-specific execution skill; specification-and-verification ability
$\theta,\ \mu,\ \tau$	latent productivity match; posterior mean; posterior precision
ρ	CARA risk aversion
$A,\ z(A)$	model capability; task-level agent competence
$\gamma_\ell,\ r_1$	gating friction on the AI mode; usage fee (both harmful)
$\lambda,\ \lambda^*$	delegated execution share; its optimum (11)
$\kappa_1,\ \varphi,\ X$	per-unit review cost; expected incident loss per delegated unit; net delegated quality (9)
$T^m,\ Z,\ \upsilon(Q),\ \Upsilon(Q)$	mode thresholds; activation indicator; inverse demand (17); gross value (18)
π_c	probability a candidate is correct
$x,\ p^{\mathrm{sv}}(x),\ p_0,\ \underline{p}$	self-verification intensity; self-pass probability of incorrect work (3); its level at $x=0$; correlated-error floor
$g,\ \hat{e},\ e$	independent-verification coverage; review standard; realized effort per reviewed candidate
$d(e),\ q^{\mathrm{FR}}(e),\ R^G$	detection function; false-rejection probability; guardrail reliability (21)
q^{FA}	false-acceptance probability (19)
$n,\ \zeta,\ C,\ C^s,\ S(x)$	agent count; candidates per agent; candidate flow; survivor flow; survivor share (22)
$Q,\ W,\ \mu_Q,\ \mu_W$	validated output; escaped defects; per-agent rates (25)–(26)
L	incident loss per escaped defect
$H^V,\ w$	human verification capacity (hours); reviewer wage
$c_A,\ c_x$	compute per candidate; self-verification cost per unit intensity
$K^G, K^C, K^F, K^R;\ r_m$	guardrail, codification, filesystem-, runtime-isolation capital; rental rates
$m_F,\ m_R$	isolation slots required per agent
$\bar{n}^V,\ \bar{n}^I,\ \bar{n}^D,\ n^u,\ n^*$	verification, isolation, demand ceilings; interior scale; implementable scale (29)
$\bar{v},\ \varsigma,\ \bar{D},\ \vartheta$	box-demand value; linear-demand slope; backlog; task-discovery rate
π_1	per-agent margin (28)
$\beta,\ \Phi_{\ell+1}$	discount factor; one-time installation cost of the stage transition
$h,\ \delta_G$	guardrail audit rate; guardrail decay
ϱ	regulatory/legacy burden (Appendix I)
$\ell,\ \sigma_\ell$	stage; stage configuration (Definition 5)

References

[1] Yash Thakker. "Boris Cherny's Steps of AI Adoption: Claude Code's 0–4 Maturity Model." explainx.ai Blog, July 17, 2026. https://explainx.ai/blog/boris-cherny-steps-ai-adoption-claude-code-july-2026

[2] Boris Cherny. "Steps of AI Adoption," LinkedIn post and public comment discussion, July 17, 2026.

[3] Anthropic. "How Claude Code is used in practice." Anthropic Research, June 16, 2026. https://www.anthropic.com/research/claude-code-expertise

[4] Anthropic. "How we built our multi-agent research system." Anthropic Engineering, June 13, 2025. https://www.anthropic.com/engineering/multi-agent-research-system

[5] Anthropic. "Multiagent orchestration." Claude Platform Docs, accessed 2026. https://platform.claude.com/docs/en/managed-agents/multiagent-orchestration

[6] Anthropic. "Hooks reference." Claude Code Docs, accessed 2026. https://code.claude.com/docs/en/hooks

[7] Thomas B. Sheridan and William L. Verplank. "Human and Computer Control of Undersea Teleoperators." Technical Report, Man-Machine Systems Laboratory, MIT, 1978.

[8] 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*, 30(3):286–297, 2000.

[9] Daron Acemoglu and Pascual Restrepo. "Automation and New Tasks: How Technology Displaces and Reinstates Labor." *Journal of Economic Perspectives*, 33(2):3–30, 2019.

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

[11] Luis Garicano. "Hierarchies and the Organization of Knowledge in Production." *Journal of Political Economy*, 108(5):874–904, 2000.

[12] Avinash K. Dixit and Robert S. Pindyck. *Investment under Uncertainty*. Princeton University Press, 1994.