A Online Appendix: A Reduced-Form Model

This appendix states the reduced-form model that underlies the article. It isolates two dynamics: the structural asymmetry between what agents can execute and what humans can verify, and alignment as an active maintenance process. Section A.6 gives the observability lever its organizational reading at the level of jobs, Section A.7 reads the system as a whole, and Section A.8 maps each claim in the article to the object or result that carries it.

A.1 Primitives and Normalizations

Category	Variable	Description
Time and tasks	$t \geq 0$	Continuous time ($\dot{x} \equiv dx/dt$)
	$i \in [0,1]$	Continuum of tasks, ordered by automation burden
Time allocation	T_m	Measurable work
	T_{nm}	Non-measurable work (steering and verification)
	T_{sim}	Synthetic practice and learning by doing
State variables	$S_{nm}(t) \geq 0$	Human experience stock
	$\tau(t) \in [0,1]$	Alignment
Deployment	$L_a(t) \geq 0$	Deployed agentic labor
Measurability	m_A, m_H	Automation frontier and human verification frontier
	$s_v,\ m_A - s_v$	Verifiable deployed mass and unverified deployed mass
	$\Delta m,\ \Delta m_+$	Measurability gap and its positive part
Risk	$X_A(t)$	Trojan Horse externality (leak flow)
Primitives	$w, B > 0$	Substitution wage (w) and verification budget (B)
	$K_C(t) > 0$	Effective automation scale
	$t_{fb}(i) \geq 0$	Feedback latency (liability horizon) of task i
Parameters	$d > 0$	Experience depreciation rate
	$\eta > 0$	Drift sensitivity
	$\kappa_{\text{corr}} \gg 1$	Correlation penalty under AI-verified AI
	$\alpha \in (0,1)$	Share of measurable capacity in effective labor
	$\phi \in (0,1)$	Output elasticity of general capital
	$\xi > 0$	Computational intensity of automated checks
	$\overline{X} > 0$	Risk budget (allowable leak flow)

Modeling notes. In the full model the cost to automate depends on task entropy and on the knowledge stocks, and the effective automation scale is $K_C^{eff}(t) = \tilde{K}_C(t)\,(A + K_{IP}(t))$, the product of compute and reusable public and proprietary knowledge. Here we re-index tasks by normalized automation burden and collapse these components into a single primitive $K_C(t)$. We treat w, B, and the path

of $K_C(t)$ as primitives and do not solve for general equilibrium; in the full model w rises with the scarcity of experienced verifiers and B is endogenized by liability and insurance.

A.2 Static Measurability Geometry

Time allocation. Output is constrained by finite human time across three rivalrous activities (theoretical education is folded into T_{sim}):

$$T_m + T_{nm} + T_{sim} \leq 1. \tag{1}$$

Cost to automate.

$$c_A(i) = \frac{i}{K_C}. \tag{2}$$

Higher i corresponds to a harder-to-automate task at a given K_C.

Cost to verify.

$$c_H(i) = w \cdot \frac{t_{fb}(i)}{S_{nm}}. \tag{3}$$

Feedback latency $t_{fb}(i)$ captures the "milliseconds versus years" dispersion in verification horizons. The budget condition can be rewritten as

$$c_H(i) < B \quad \Longleftrightarrow \quad S_{nm} > \frac{w}{B}\, t_{fb}(i),$$

which states the two routes to expanding verification: raise experience S_{nm} (augmentation and T_{sim}) or compress latency $t_{fb}(i)$ (observability and provenance). Liability and insurance operate on B.

Frontiers, gap, and verifiable share.

$$m_A \equiv \int_0^1 \mathbb{1}[c_A(i) < w]\, di, \qquad m_H \equiv \int_0^1 \mathbb{1}[c_H(i) < B]\, di, \qquad \Delta m \equiv m_A - m_H. \tag{4}$$

$$s_v \equiv \int_0^1 \mathbb{1}[c_A(i) < w \ \cap \ c_H(i) < B]\ di. \tag{5}$$

Tasks with $c_A(i) < w$ and $c_H(i) < B$ form the safe industrial zone, so s_v is the mass of deployed agentic activity that is both automatable and affordably verifiable. The unverified deployed mass is

$$m_A - s_v = \int_0^1 \mathbb{1}\{c_A(i) < w \ \wedge \ c_H(i) \geq B\}\, di,$$

the runaway risk zone. Because Δm is a difference between the measures of two sets, it can be negative when the artisan region is large; what drives risk is $m_A - s_v$. In the limit $K_C \to \infty$, $m_A \to 1$ and $s_v \to m_H$, so $\Delta m \to 1 - m_H = m_A - s_v$. We therefore use Δm (and $\Delta m_+ \equiv \max\{\Delta m, 0\}$) as the reduced-form pressure gauge for drift. This is a maintained approximation for drift, not an identity with unverified deployment outside the compute limit: $m_A - s_v = \Delta m + (m_H - s_v)$ can be positive even when $\Delta m \leq 0$.

Measurable capacity and the leak. Only verified agentic output contributes to measurable capacity:

$$L_m = T_m + s_v\, L_a. \tag{6}$$

Non-measurable capacity is $L_{nm} = S_{nm} T_{nm}$. As in the full paper, $Y = A K_G^{\phi} L_E^{1-\phi}$ with $L_E = L_{nm}^{1-\alpha} L_m^{\alpha}$, where A is public knowledge and K_G is general capital. Thus $\partial Y / \partial L_m = (1-\phi)\alpha Y / L_m$. The unverified, misaligned remainder leaks as the Trojan Horse externality,

$$X_A = (1-\tau)\,(m_A - s_v)\, L_a. \tag{7}$$

Volume is $(m_A - s_v)L_a$, the agentic labor operating in the blind spots; severity is $(1 - \tau)$, the degree of drift from intent. The leak enters capital accumulation as a predator on the surplus, $\dot{K}_G = Y - C_Y - X_A - \delta_K K_G$, competing with consumption and investment.

Proposition A.1 (Structural asymmetry of measurability). *Agent measurability m_A scales with K_C, while human measurability m_H is bounded by human experience S_{nm}.*

(1) *As $K_C \to \infty$, $m_A \to 1$.*
(2) *Holding institutions fixed, m_H increases in S_{nm} and B and decreases in w and t_{fb}, but is limited by the dynamics of S_{nm} in* (8).
(3) *Holding the automatable set fixed, a pointwise rise in c_H or a fall in B weakly reduces s_v. If the automatable set expands while the verifiable set contracts, the unverified mass $m_A - s_v$ weakly rises, although s_v may move in either direction.*

Proof. From (2), $m_A = \int_0^1 \mathbb{1}[i/K_C < w]\, di = \min\{1, wK_C\}$, so $\partial m_A/\partial K_C \geq 0$ and $\lim_{K_C \to \infty} m_A = 1$. From (3), $c_H(i)$ is pointwise decreasing in S_{nm} and increasing in w and $t_{fb}(i)$, so the set $\{i : c_H(i) < B\}$ expands in S_{nm} and B and m_H is weakly increasing in those arguments. For (3), write $M_A = \{i : c_A(i) < w\}$ and $M_H = \{i : c_H(i) < B\}$. With M_A fixed, contraction of M_H shrinks $M_A \cap M_H$, so s_v weakly falls. If $M_A \subseteq M_A'$ and $M_H' \subseteq M_H$, then $M_A \setminus M_H \subseteq M_A' \setminus M_H'$, so $m_A - s_v$ weakly rises; the two intersections need not be ordered.

□

Remark 1 (Human augmentation as effective verification bandwidth). *Augmentation raises the effectiveness of scarce oversight time by raising effective S_{nm} per unit time and/or reducing effective $t_{fb}(i)$ through observability. Both shift $c_H(i)$ downward pointwise, expanding m_H and s_v and reducing Δm. Verification-grade proprietary data (audited traces, incident registries, outcome archives, provenance logs) belongs to this channel: it makes claims and failures quickly adjudicable, which lowers effective $t_{fb}(i)$. So does oversight tooling (observability, interpretability, continuous monitoring), which raises the verification bandwidth of each hour of T_{nm}.*

Remark 2 (Capability does not close the gap). *The verification frontier $m_H = \int_0^1 \mathbb{1}[w\, t_{fb}(i)/S_{nm} < B]\, di$ depends only on $(w, B, t_{fb}(\cdot), S_{nm})$ and on nothing about the agent. Hence, for any human-side state, $\partial m_H/\partial K_C = 0$ while $\partial m_A/\partial K_C \geq 0$, so $\partial \Delta m/\partial K_C \geq 0$, and in the limit $\Delta m \to 1 - m_H$ is pinned by the human side alone. A more capable executor facing the same verifiers faces a larger gap, not a smaller one; this is the sense in which the constraint survives the transition from AGI to any level of superintelligence. Compute can lower effective $t_{fb}(i)$ through observability tooling, but that is a measurement investment (channel (iii) of Proposition A.4), not a capability one, and it is the only route by which the machine side touches m_H.*

A.3 Dynamics: Human Experience and Alignment

Dynamic Law 1 (the Missing Junior Loop). The experience stock is maintained through measurable execution or synthetic practice:

$$\dot{S}_{nm} = T_m + T_{sim} - d\, S_{nm}. \tag{8}$$

For constant (T_m, T_{sim}), a steady-state experience stock at or above $\underline{S}_{nm} > 0$ requires inflow to cover depreciation at the floor:

$$T_m + T_{sim} \geq d\, \underline{S}_{nm} \quad \Longleftrightarrow \quad T_{sim} \;\geq\; \max\{0,\, d\, \underline{S}_{nm} - T_m\}.$$

The maintenance rule $T_{sim}(t) = \max\{0,\, d\, \underline{S}_{nm} - T_m(t)\}$ supplies the missing inflow, subject to the time budget (1). It gives $\dot{S}_{nm} \geq 0$ at the floor and moves a stock below it toward the target. Figure 3 illustrates stepped increases in practice. Coupled with $S_{nm} > \frac{w}{B} t_{fb}(i)$, the floor is the critical mass

of experience needed to keep the long-horizon tail of tasks, those with $t_{fb}(i) < \frac{B}{w}\underline{S}_{nm}$, inside the verifiable set. To close the feedback loop we impose a reduced-form substitution condition: holding (T_{nm}, T_{sim}) fixed, measurable human execution weakly falls as agent measurability expands, $\partial T_m / \partial m_A \leq 0$.

Proposition A.2 (The Missing Junior Loop). *Under $\partial T_m/\partial m_A \leq 0$, automation of routine execution erodes verification capacity by severing the experience accumulation channel.*

(1) *For fixed (T_m, T_{sim}), (8) has a unique, globally stable steady state $S^{\star}_{nm} = (T_m + T_{sim})/d$.*
(2) *An expansion of m_A (via K_C) weakly reduces T_m; unless offset by T_{sim}, $S^{\star}_{nm}$ weakly falls, $c_H(i)$ rises pointwise, and m_H weakly falls. Holding the automatable set fixed at its new level, this loss of verification capacity weakly contracts s_v. Across the combined expansion of automation and contraction of verification, $m_A - s_v$ weakly rises, while s_v can move either way.*

Proof. Setting $\dot{S}_{nm} = 0$ gives $S^{\star}_{nm} = (T_m + T_{sim})/d$, and $S_{nm}(t) = S^{\star}_{nm} + (S_{nm}(0) - S^{\star}_{nm})e^{-dt}$ establishes global stability. By the substitution assumption, higher m_A weakly lowers T_m and hence $S^{\star}_{nm}$. Since $c_H(i) = w\, t_{fb}(i)/S_{nm}$ is strictly decreasing in S_{nm}, a fall in S_{nm} raises $c_H(i)$ for all i, shrinking $\{i : c_H(i) < B\}$ by set inclusion, so m_H weakly falls. For a fixed automatable set, the intersection defining s_v shrinks; allowing automation to expand, Proposition A.1(3) gives the stated result for unverified mass. □

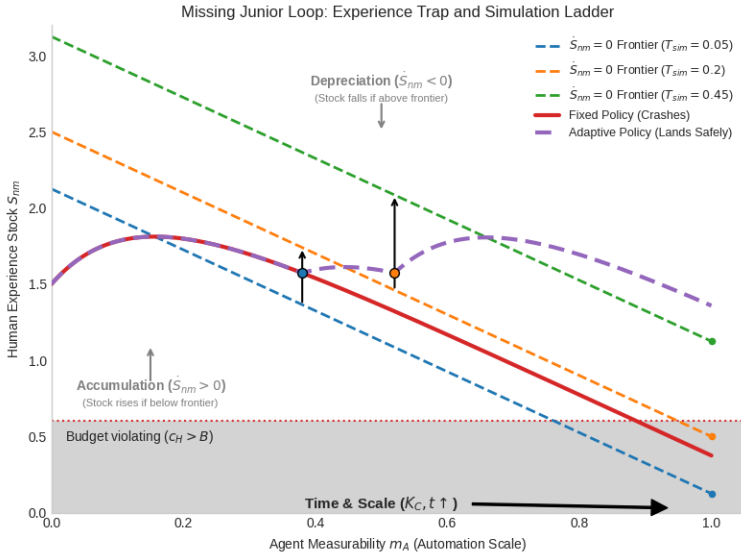


Fig. 3. **Experience trap and simulation ladder.** Nullcline diagram for the experience stock under (8). The dashed curves plot $\dot{S}_{nm} = 0$ for increasing levels of synthetic practice T_{sim}, using $T_m(m_A) = T_{m0}(1 - m_A)$ so that the steady state $S^{\star}_{nm} = (T_{m0}(1 - m_A) + T_{sim})/d$ falls as automation rises. The shaded region marks the budget-violating zone for high-stakes verification ($c_H > B$). Under a fixed policy with low T_{sim} (solid red), the trajectory crashes into the unverified zone. Under an adaptive policy (dashed purple), T_{sim} is stepped up at trigger points, shifting the nullcline upward and landing S_{nm} above the threshold despite the decline in T_m.

Remark 3 (The Codifier's Curse). *In the full model the effective automation scale is $K_C(t)\,(A + K_{IP})$, and the proprietary stock accumulates partly through tacit extraction: verification and steering*

generate machine-readable traces (labels, redlines, code reviews, audit logs), so $\dot{K}_{IP} \propto T_{nm}$. With $\partial c_A(i)/\partial K_{IP} < 0$, any task subjected to steering experiences $\dot{c}_A(i) = (\partial c_A(i)/\partial K_{IP})\,\dot{K}_{IP} \le 0$, so $\{i : c_A(i) < w\}$ expands and $\partial m_A/\partial T_{nm} \ge 0$ holding physical compute fixed. The same scarce verification effort that raises m_H today raises m_A tomorrow. In the full model the expert's wage is $w(S_{nm})$ with a scarcity premium, which is why top verifiers gain value in the transition even as their traces erode it; experts supply T_{nm} only while the current wage exceeds the discounted induced erosion, and because they cannot coordinate to withhold data, extraction proceeds at the maximum rate.

Dynamic Law 2 (Alignment as Maintenance).

$$\dot{\tau} = (1-\tau)\,T_{nm} \;-\; \tau \cdot \eta \cdot \Delta m_{+}, \qquad \Delta m_{+} \equiv \max\{\Delta m, 0\}. \tag{9}$$

The first term is active maintenance with diminishing returns as $\tau \to 1$; the second is structural drift pressure driven by the positive gap. Here T_{nm} bundles all steering and verification effort that relies on non-measurable human time, and η is shaped by base-alignment and graceful-degradation design choices. Augmentation enters by raising the effective bandwidth of T_{nm} or lowering t_{fb}, which shrinks Δm. Realized exposure is $m_A - s_v$; inherited base alignment is the initial condition $\tau(0)$.

Proposition A.3 (Alignment as maintenance). *For constant T_{nm} and Δm, alignment converges to*

$$\tau^{\star} = \frac{T_{nm}}{T_{nm} + \eta\,\Delta m_{+}}.$$

If oversight is abandoned ($T_{nm} = 0$) and $\Delta m_{+} > 0$, alignment decays exponentially, $\tau(t) = \tau(0)e^{-\eta\,\Delta m_{+}\,t}$. The interval $[0,1]$ is forward invariant.

Proof. With constant parameters, (9) is the linear ODE $\dot{\tau} = T_{nm} - \tau(T_{nm} + \eta\,\Delta m_{+})$; setting $\dot{\tau} = 0$ gives $\tau^{\star}$. With $T_{nm} = 0$ the equation reduces to $\dot{\tau} = -\eta\,\Delta m_{+}\,\tau$, whose solution is the stated exponential. At $\tau = 0$, $\dot{\tau} = T_{nm} \ge 0$; at $\tau = 1$, $\dot{\tau} = -\eta\,\Delta m_{+} \le 0$; the field points inward at both boundaries. □

AI verifying AI. If firms substitute compute for experience in verification, the cost of obtaining a check becomes

$$c_H^{*}(i) = \min\left\{ w\frac{t_{fb}(i)}{S_{nm}},\; \frac{\xi}{K_C} \right\},$$

with ξ the computational intensity of verification. As K_C scales, more tasks clear this cost threshold, but cheaper checks alone do not establish reliable coverage; s_v retains its baseline definition in (5). To describe an adverse scenario in which cheap, correlated checks substitute for effective human review, assume that drift sensitivity becomes

$$\eta_{\text{eff}} = \begin{cases} \eta & \text{if human verified,} \\ \eta \cdot \kappa_{\text{corr}} & \text{if AI verified, with } \kappa_{\text{corr}} \gg 1. \end{cases}$$

The correlation multiplier is a maintained assumption for this adverse scenario. For fixed $T_{nm} > 0$ and $\Delta m_{+} > 0$, the higher drift sensitivity lowers the steady-state alignment in Proposition A.3, even as checking costs fall.

The risk budget. Treating unverified throughput as latent debt amounts to the constraint $X_A(t) \le \overline{X}$, which caps deployment at

$$L_a(t) \;\le\; \frac{\overline{X}}{(1-\tau(t))\,(m_A(t) - s_v(t))} \qquad \text{when } (1-\tau(t))(m_A(t) - s_v(t)) > 0.$$

Autonomy and scale are conditioned on auditability and insurability.

A.4 Deployment Incentives and the Tipping Point

A deployer chooses L_a and the verification budget B to maximize $\Pi = p\,Y(L_a, s_v(B)) - C(B) - r_C K_C - \ell\, X_A$, where $C(B)$ is verification expenditure and $\ell \geq 0$ is the priced liability wedge. Taking L_a and the other state variables as given, the verification subproblem is

$$B^\star \in \arg\min_{B \geq 0} \Big\{ C(B) + \ell(1-\tau)\big(m_A - s_v(B)\big)L_a - pY(L_a, s_v(B)) \Big\},$$

so a differentiable interior optimum satisfies $C'(B^\star) = \big[p(1-\phi)\alpha(Y/L_m)L_a + \ell(1-\tau)L_a\big]s_v'(B^\star)$. Verification earns a productive return as well as avoiding priced losses. With $s_v'(B) \geq 0$ and a strict second-order condition, the local comparative statics are $\partial B^\star/\partial \ell \geq 0$, $\partial B^\star/\partial L_a \geq 0$, and $\partial B^\star/\partial \tau \leq 0$. As $\ell \to 0$, verification is determined by its private productive return and need not collapse to its floor.

Verification also has public good components: shared standards, evaluation tools, and incident knowledge can benefit deployers who did not fund them. These spillovers can make private investment fall short of the social optimum. We retain this economic interpretation without asserting a Nash condition for the continuum of deployers. Liability and insurance can also increase private incentives to verify by raising ℓ.

Finally, for $L_m > 0$ the net marginal product of an agent is $\mathrm{MP}_{net} = (1-\phi)\alpha(Y/L_m)s_v - (1-\tau)(m_A - s_v)$. When $m_A - s_v > 0$, it is zero at

$$\tau_{\text{crit}} = 1 - \frac{(1-\phi)\alpha\, Y\, s_v}{L_m\,(m_A - s_v)}.$$

At a given state, an additional agent contributes more than it leaks when $\tau > \tau_{\text{crit}}$ and is net-extractive when $\tau < \tau_{\text{crit}}$. A feasible sign crossing requires the threshold to lie in $[0, 1]$. This marginal-product condition does not establish distinct equilibria or a bifurcation; the hollow and augmented economies are interpretations of alternative regimes.

A.5 Governance as Primitive Shifts

Proposition A.4 (Governance as primitive shifts). *Governance operates by shifting primitives that expand the verification frontier.*

(1) ***(Liability and enforcement)*** *For any $B_2 > B_1$: $m_H(B_2) \geq m_H(B_1)$, $s_v(B_2) \geq s_v(B_1)$, $\Delta m(B_2) \leq \Delta m(B_1)$. Interventions that force deployers to internalize X_A raise $B^\star$ endogenously.*

(2) ***(Learning and simulation)*** *For any $T_{sim,2} > T_{sim,1}$, holding T_m fixed, $S_{nm}^\star$ rises; holding (w, t_{fb}) and the automatable set fixed, m_H and s_v weakly rise, Δm weakly falls, and $\tau^\star$ weakly rises for fixed T_{nm}.*

(3) ***(Observability and human augmentation)*** *Any intervention that lowers $t_{fb}(i)$ or raises effective S_{nm} per unit time lowers $c_H(i)$ pointwise, expanding m_H and, for a fixed automatable set, s_v, reducing Δm, and weakly raising $\tau^\star$ for any given T_{nm}.*

Proof. (1) Let $M_H(B) \equiv \{i : c_H(i) < B\}$. If $B_2 > B_1$ then $M_H(B_1) \subseteq M_H(B_2)$, so $m_H(B_2) \geq m_H(B_1)$; with $M_A \equiv \{i : c_A(i) < w\}$ fixed, $s_v(B) = \mathrm{meas}(M_A \cap M_H(B))$ is weakly increasing, and $\Delta m(B) = m_A - m_H(B)$ weakly decreasing. (2) From (8), $S_{nm}^\star = (T_m + T_{sim})/d$ is increasing in T_{sim}; since $c_H(i)$ is decreasing in S_{nm}, $M_H(B)$ expands, so m_H and s_v weakly rise and Δm weakly falls; substituting into $\tau^\star = T_{nm}/(T_{nm} + \eta\,\Delta m_+)$ gives a weakly higher steady state. (3) $c_H(i) = w\,t_{fb}(i)/S_{nm}$ is increasing in $t_{fb}(i)$ and decreasing in S_{nm}, so the stated interventions lower it pointwise; the rest follows as in (2) with K_C held fixed. □

A.6 Jobs as Bundles: Metrics Capital at the Interface

The geometry above is defined over tasks, but firms hire and reorganize jobs, which are bundles of subtasks [22]. As the cost of automating some subtasks falls, the firm faces pressure to unbundle: agents take the automatable components and humans keep the residual. The question is not only whether a subtask is automatable but whether the reorganized bundle is still verifiable. When one actor executes a workflow end to end, intermediate state, tacit assumptions, and accountability stay inside the bundle; splitting it pushes that context across an organizational interface and raises effective feedback latency. Metrics capital (audit trails, evaluation harnesses, provenance receipts, reconciliation checks) is the firm-level investment that rebuilds the lost context. This subsection is the organizational reading of the observability lever in Proposition A.4(3), not a separate framework.

Partition the task interval into job cells $\{J_j\}$ with masses $\omega_j > 0$, $\sum_j \omega_j = 1$, each discretized into K_j equal-mass subtasks with baseline automation cost $c_{A,jk}$ and baseline latency $t_{fb,jk}$. Let $n_j \in [0,1]$ be the degree of AI-enabled unbundling of job j and $M_j \geq 0$ the firm's metrics capital. Effective latency and verification cost are

$$\tilde{t}_{fb,jk}(n_j, M_j) = t_{fb,jk}\,\frac{1+\beta_j n_j}{1+\chi_j M_j}, \qquad c_{H,jk}(n_j, M_j) = w\,\frac{\tilde{t}_{fb,jk}(n_j, M_j)}{S_{nm}}, \tag{10}$$

where $\beta_j \geq 0$ measures how much splitting destroys verification-relevant context and $\chi_j > 0$ how effectively measurement rebuilds it. The within-job automatable and verifiable shares are

$$m_{A,j} = \frac{1}{K_j}\sum_{k=1}^{K_j} \mathbb{1}\{c_{A,jk} < w\}, \qquad s_{v,j}(n_j, M_j) = \frac{1}{K_j}\sum_{k=1}^{K_j} \mathbb{1}\{c_{A,jk} < w \wedge c_{H,jk}(n_j, M_j) < B\}, \tag{11}$$

and the within-job unverified shortfall is $\Gamma_j(n_j, M_j) \equiv m_{A,j} - s_{v,j}(n_j, M_j)$, the share of subtasks in bundle j that are cheap to automate but not affordably verifiable after reorganization. Aggregating recovers the core objects without a new state variable:

$$m_A = \sum_j \omega_j\, m_{A,j}, \qquad s_v = \sum_j \omega_j\, s_{v,j}, \qquad m_A - s_v = \sum_j \omega_j\, \Gamma_j. \tag{12}$$

The firm chooses (n_j, M_j) to maximize $G_j(n_j; K_C) - \kappa_j(M_j) - \ell_j L_{a,j}\,\Gamma_j(n_j, M_j)$, where G_j is the gross gain from peeling automatable subtasks out of the bundle, with $\partial G_j/\partial n_j > 0$ and $\partial^2 G_j/\partial n_j \partial K_C > 0$ so that rising capability raises the private return to unbundling, κ_j is convex, and $\ell_j \geq 0$ is the liability wedge of Appendix A.4. Because Γ_j is built from threshold conditions, it is piecewise constant in (n_j, M_j), so we characterize the comparative statics through monotonicity rather than first-order conditions.

Proposition A.5 (Bundles, splitting, and endogenous measurement). *Fix job j and hold $\{c_{A,jk}\}_{k=1}^{K_j}$, S_{nm}, and B fixed.*

(1) *$\tilde{t}_{fb,jk}$ and $c_{H,jk}$ are weakly increasing in n_j and weakly decreasing in M_j for every k.*
(2) *$s_{v,j}$ is weakly decreasing in n_j and weakly increasing in M_j.*
(3) *Γ_j is weakly increasing in n_j and weakly decreasing in M_j.*

Proof. From (10), $\partial\tilde{t}_{fb,jk}/\partial n_j = t_{fb,jk}\,\beta_j/(1+\chi_j M_j) \geq 0$ and $\partial\tilde{t}_{fb,jk}/\partial M_j = -t_{fb,jk}\,\chi_j(1+\beta_j n_j)/(1+\chi_j M_j)^2 \leq 0$, and $c_{H,jk}$ inherits both signs, which gives (1). The safe set $\{k : c_{A,jk} < w,\ c_{H,jk}(n_j, M_j) < B\}$ therefore contracts weakly in n_j and expands weakly in M_j; since $s_{v,j}$ is its measure, (2) follows. With $m_{A,j}$ fixed, $\Gamma_j = m_{A,j} - s_{v,j}$ moves in the opposite direction, which is (3). □

Remark 4 (Bottleneck bundles). *Γ_j is the right statistic when deployability depends on the average shortfall across subtasks. In O-ring bundles [29], one unverifiable subtask blocks the whole*

reorganization and the relevant object is $\Gamma_j^{\max} \equiv \max_k \mathbb{1}\{c_{A,jk} < w \wedge c_{H,jk} \geq B\}$. *The monotonicity in Proposition A.5 is unchanged, but the return to metrics capital becomes threshold-like: lowering verification cost on the other subtasks does nothing visible until the critical one crosses below B.*

The 655-firm study in the text is the mechanism at the scale of one job. Peeling code generation out of the engineer's bundle moved the context a reviewer needs across an interface ($n_j \uparrow$), the elapsed time from submission to merge rose 56 percent, consistent with greater review congestion ($\Gamma_j \uparrow$), and the review tooling firms adopted alongside is M_j. Value shifts toward organizations that can rebuild verification fast enough to keep the reorganized bundle inside the verifiable frontier.

A.7 Reading the Model

The skeleton is one coupled system. Compute scaling raises agent measurability ($K_C \uparrow \Rightarrow m_A \uparrow$); competitive substitution crowds out routine human execution ($T_m \downarrow$); the experience stock decays unless synthetic practice offsets it ($S_{nm} \downarrow$ absent $T_{sim} \uparrow$); verification costs rise and the human verification frontier contracts ($c_H \uparrow$, $m_H \downarrow$), while s_v may move either way as automation expands; and, once positive, the widening gap creates drift pressure absent continuous steering ($\Delta m \uparrow \Rightarrow \tau \downarrow$ unless $T_{nm} \uparrow$). The hazard is the unverified, misaligned mass $(1-\tau)(m_A - s_v)$. The institutional objective is to maximize verified deployment $s_v L_a$ while pricing the leak X_A: simulation to rebuild S_{nm}, observability and metrics capital to compress t_{fb}, liability to raise B.

Two boundaries follow. Synthetic practice rebuilds experience only where the simulator captures the relevant reality; where it does not, new ground truth has to be generated by directed experimentation. The threshold τ_{crit} in Appendix A.4 distinguishes positive from negative net marginal contributions at a given state. The hollow and augmented economies describe alternative regimes; the reduced form does not establish them as separate equilibria or prove a bifurcation between them. Sustained human oversight can support steady-state alignment under Proposition A.3; maintaining it requires continuing effort.

A.8 Crosswalk: Storyline to Model

Claim in the article	Model object	Result
What can be measured can be automated	$c_A(i) = i/K_C$; $m_A = \min\{1, wK_C\} \to 1$ as compute and knowledge scale	Prop. A.1(1)
Context and intent are tacit, recorded in experience through execution and out-of-distribution exposure	S_{nm} as a stock; $\dot{S}_{nm} = T_m + T_{sim} - dS_{nm}$; $L_{nm} = S_{nm} \cdot T_{nm}$	Law 1; Prop. A.2(1)
Output that automation cannot fully verify needs a human on the other end	$c_H(i) = w\, t_{fb}(i)/S_{nm}$; m_H; s_v; $L_m = T_m + s_v L_a$	Eqs. (3)–(6); Prop. A.1(3)
Verification is the bottleneck through AGI and remains so under superintelligence	m_H depends only on (w, B, t_{fb}, S_{nm}); $\partial\Delta m/\partial K_C \geq 0$; $\Delta m \to 1 - m_H$	Remark 2
The gap is a source of risk and private incentives are perverse	$X_A = (1-\tau)(m_A - s_v)L_a$; priced liability wedge ℓ	Eq. (7); App. A.4 (public-good interpretation)

Claim in the article	Model object	Result
Top verifiers gain value temporarily; their traces drive the codifier's curse	scarcity premium $w(S_{nm})$; $\dot{K}_{IP} \propto T_{nm}$; $\partial m_A / \partial T_{nm} \geq 0$	Remark 3
Juniors face the missing junior loop because their knowledge overlaps with the model's	$\partial T_m / \partial m_A \leq 0$ lowers $S^{\star}_{nm} = (T_m + T_{sim})/d$	Prop. A.2(2)
Answer to the curse: elevate humans to directing agents and holding intent	T_{nm} as steering in Law 2; augmentation raises effective T_{nm}	Eq. (9); Prop. A.4(3)
Answer to the junior loop: flight simulators for work and AI-driven discovery of new ground truth	$T_{sim} \geq d\,\underline{S}_{nm} - T_m$; effective scale $\tilde{K}_C(A + K_{IP})$	Law 1; Prop. A.4(2)
The gap drives alignment drift; correlated AI checks can create false confidence	$\dot{\tau} = (1-\tau)T_{nm} - \tau\eta\Delta m_+$; $\tau^{\star}$; assumed $\eta_{\text{eff}} = \eta\kappa_{\text{corr}}$	Prop. A.3; adverse scenario in App. A.3
Only verification infrastructure and human augmentation close the gap	$t_{fb} \downarrow$, effective $S_{nm} \uparrow \Rightarrow m_H \uparrow$, $s_v \uparrow$, $\Delta m \downarrow$	Prop. A.4(3); Remark 1
Verification infrastructure as a moat: rebuilding verification at the interface after AI unbundles a job	$\tilde{t}_{fb,jk} = t_{fb,jk}(1+\beta_j n_j)/(1+\chi_j M_j)$; Γ_j; $m_A - s_v = \sum_j \omega_j \Gamma_j$	Prop. A.5; Remark 4
Policy: liability-as-a-service and a market for insuring agentic outcomes, with telemetry	$B^{\star}$ increasing in ℓ; risk budget $L_a \leq \overline{X}/[(1-\tau)(m_A - s_v)]$; τ_{crit}	Prop. A.4(1); App. A.4
Moats: top talent, ground truth, the unmeasurable	S_{nm} and $w(S_{nm})$; verification-grade K_{IP} lowering c_H; tasks with $c_A \geq w$ and $c_H \geq B$ (tacit zone)	Fig. 1; Remarks 1, 3; Prop. A.5