The Task Frontier
Horizon, Forgiveness, and the Direction of AI Innovation

July 2026

Abstract

Cheap AI systems commoditize existing tasks within months, yet frontier laboratories keep raising their spending. This paper explains the coexistence with a theory in which the state of AI is a two-dimensional task frontier and imitation is directionally asymmetric. Tasks are O-ring production with retries: n serial stages, each permitting ℓ retries. This one primitive *delivers* the two dimensions rather than assuming them: required reliability rises log-linearly in horizon with slope $1/(1+\ell)$, task success converges to a Gumbel kernel, and a scalar quality ladder is exactly the special case of homogeneous forgiveness. Price competition task by task implies that frontier profit equals the boundary-value integral over the capability gap, so rents survive only where followers cannot follow. Imitation is fast along the horizon dimension—a plan can be copied from one demonstration—and slow along the reliability dimension—a failure rate must be estimated from many. Equilibrium investment therefore rotates from horizon toward reliability as followers catch up: the market races along the moat, not along value. The resulting direction of innovation is distorted relative to the planner's by the imitation asymmetry, an observable object. Liability rules and public verification infrastructure both raise reliability investment but move market concentration in opposite directions. Aggregate evidence—horizon doubling times, the 50%/80%-horizon gap, and asymmetric open-weight catch-up lags—disciplines the model's key objects.

JEL: O31, O33, L13, L15, D24.

1 Introduction

Two facts about the market for artificial intelligence sit awkwardly together. The price of any fixed, demonstrated capability collapses: across public benchmarks, the cost of reaching a given score has fallen by one to two orders of magnitude per year, and open-weight models reproduce frontier scores within months. At the same time, the laboratories that own the frontier keep raising their investment. If artificial intelligence were a single quality index, this

pattern would be a puzzle: when imitation is nearly immediate, the return to advancing the index should be competed away, and frontier spending should fall, not rise.

This paper resolves the puzzle with two ingredients. First, the state of AI is not a scalar; it is a frontier in a two-dimensional task space. One dimension is *horizon*: the serial depth of the actions a task requires. The other is *forgiveness*: how much the task permits errors to be observed, retried, and rolled back before they become losses. Second, imitation is not a scalar hazard; it is *directional*. The horizon dimension of capability is embodied in observable plans, scaffolds, and reasoning traces, and can be copied from demonstrations. The reliability dimension is a tail property: certifying that a system fails less than once in ten thousand attempts requires on the order of ten thousand independent trials, and the unforgiving tasks where such reliability matters are exactly the tasks that generate few observable trials. Rents are competed away along the dimension where imitation is fast and survive along the dimension where it is slow. Frontier investment follows the rents: as followers close the horizon gap, equilibrium investment rotates toward the reliability boundary. The market races along the moat, not along value.

The paper's first contribution is to derive the two-dimensional frontier from one primitive rather than to posit it. A task is O-ring production with retries: n stages must all succeed, and each stage may be attempted $1+\ell$ times (Kremer, 1993). For a system whose per-attempt failure rate declines in a reliability capability q_F, the requirement to complete the task with tolerated failure ε reduces, exactly and transparently, to a threshold

$$q_F \;\geq\; \varphi + \frac{h + \ln(1/\varepsilon)}{1+\ell}, \qquad h = \ln n,$$

up to a vanishing correction (Section 3). Horizon imposes a reliability tax—serial depth compounds residual error—and forgiveness divides that tax by $1+\ell$. Task success as a function of capability converges to a Gumbel kernel whose location is this threshold and whose scale is $1/(1+\ell)$, which microfounds the reduced-form success functions used in task-based models. The representation theorem (Theorem 1) then states exactly when the familiar scalar ladder is right: the task economy admits a one-dimensional quality representation if and only if task requirements form a chain—equivalently, if forgiveness is homogeneous across tasks. Coding is long but forgiving; a payment authorization is short but unforgiving. Once forgiveness is heterogeneous, no scalar index can order tasks by difficulty, and the direction of progress becomes an economic object rather than a redundant label.

The second contribution is to derive, rather than assume, the market payoffs that task-based theories of AI competition have treated as primitives. Firms compete in prices task by task, as in asymmetric Bertrand competition with vertical differentiation. The equilibrium allocation is the familiar sorting rule—a cheaper follower wins a task exactly when its cost advantage exceeds the remaining value gap—and the frontier firm's flow profit equals the *boundary-value*

integral over the capability gap between itself and the best alternative (Lemma 2). Saturation is then an equilibrium state, not an assumption: as the fringe closes the gap in a task region, markups there converge to zero mechanically, while total surplus stays high. Regions of large current revenue and zero frontier rent coexist with regions of small current revenue and large frontier value.

The third contribution is the imitation technology. Theorem 2 formalizes the asymmetry: a follower that observes a single successful trajectory learns the plan, so the horizon component of capability diffuses at a rate bounded away from zero; but any procedure that certifies a failure rate below ε from observed behavior requires $\Omega(1/\varepsilon)$ independent observations, and the tolerated failure rate at the fragility frontier is exponentially small in the frontier's position. The cost of imitating reliability therefore grows exponentially as the reliability frontier advances, while the cost of imitating demonstrated plans does not. Appropriability is not an institutional parameter here; it is a statistical property of tasks. Forgiving, observable, easily evaluated tasks are cheap to distill. Unforgiving tasks protect their own rents.

These three pieces assemble into a theory of the direction of innovation. In the dynamic model, a frontier laboratory invests against a competitive fringe that closes the capability gap at rate λ_H in the horizon dimension and λ_F in the reliability dimension, with $\lambda_H > \lambda_F$ by Theorem 2. The laboratory's shadow value of a unit of capability in direction i is $B_i/(r+\lambda_i)$: the boundary value of tasks unlocked in that direction, priced by the *persistence* of the gap there (Theorem 3). Investment follows persistence-adjusted boundary values. Because pushing the horizon frontier unlocks long tasks whose binding constraint is then reliability—the cross-boundary complementarity is positive by construction—the value ratio B_F/B_H rises along any horizon push, and investment rotates in finite time toward the reliability boundary. Faster imitation of the horizon dimension makes the rotation come *earlier*: imitation does not merely deter frontier investment, it steers it. Strategic competition sharpens the result: when two laboratories choose which boundary to race on, sufficiently strong imitation asymmetry makes racing on the moat dimension a dominant strategy, so competitors herd onto the same unforgiving frontier even though they split the prize there (Theorem 4).

The welfare economics is transparent and, we think, the right way to pose the policy question. A planner values a unit of direction-i capability at B_i/r: capability is permanent for society, and the fringe's catch-up is diffusion, not destruction. The market values it at $B_i/(r+\lambda_i)$. The direction of innovation is therefore distorted by the factor $(r+\lambda_F)/(r+\lambda_H)$—the market undervalues exactly the dimension that is easy to imitate—and the sign and size of the distortion are pinned by an observable, the asymmetry in imitation lags (Theorem 5). Two policies that both "promote reliable AI" then have opposite competitive effects. Liability for failures in unforgiving tasks raises the private demand price of reliability, accelerates investment in it, and *thickens* the moat: concentration rises. Public verification infrastructure—shared evaluation suites, incident registries, certification institutions—raises λ_F, diffuses reliability,

and *thins* the moat: concentration falls, at the cost of weakening the private incentive it disciplines. Safety policy and competition policy are the same policy lever pointed in different directions.

The paper is theoretical, and the discipline we impose is aggregate, in the style of quantitative theories of ideas and growth (Bloom et al., 2020). Three public facts anchor the model's objects. First, the measured horizon frontier—the task duration at which frontier systems succeed half the time—has doubled roughly every seven months (Kwa et al., 2025). Second, the same measurement contains the second dimension: horizons measured at an 80% success requirement are roughly five times shorter than at 50%. A scalar constant-hazard benchmark (Ord, 2025) makes a parameter-free prediction for this ratio, $\ln 0.5/\ln 0.8 \approx 3.1$; the observed gap of roughly five is the reliability tax the scalar model misses, and its evolution identifies the coupling parameter $1/(1+\ell)$. Third, imitation lags are measurably asymmetric: open-weight models trail frontier capability indices by three to four months, but trail frontier *autonomy* time-horizons by six to twelve months, and the gap at high reliability thresholds is wider still. Section 9 uses these objects to place the current race on the model's phase diagram.

Related literature. The paper builds on task-based models of technology (Acemoglu and Autor, 2011; Acemoglu and Restrepo, 2018) but moves the action from the worker–task substitution margin to the structure of tasks themselves, and on directed technical change (Acemoglu, 2002, 2007), from which it inherits the question "which direction does the market choose, and is it the right one?" Its answer mechanism is new: in Acemoglu (2002) direction is steered by market size and factor prices; here it is steered by the statistical structure of imitation. The closest antecedents on the direction of innovation are Bryan and Lemus (2017) and Dasgupta and Maskin (1987), where racing distorts which projects firms pursue, and Callander (2011), where innovators search an unknown landscape; Budish et al. (2015) document direction distortions empirically. Relative to this work, the direction space here is derived from a task primitive, the distortion is signed by an observable (imitation-lag asymmetry), and the dynamics deliver rotation and herding rather than a static portfolio bias. The O-ring production structure descends from Kremer (1993) and its weakest-link successors (Jones, 2011); forgiveness (retries) is the new margin, and it is what makes the task space two-dimensional. The racing and persistence-of-leadership questions connect to patent races and preemption (Gilbert and Newbery, 1982; Reinganum, 1983) and to step-by-step innovation and escape competition (Aghion et al., 2001); the rotation result is an escape-competition force acting on the *direction* rather than the rate of innovation. The appropriability mechanism formalizes, for AI, the classic observations of Arrow (1962), Teece (1986), and Anton and Yao (1994, 2002): what cannot be disclosed or copied earns the rents. Multidimensional quality competition goes back to Mussa and Rosen (1978) and Shaked and Sutton (1982); the contribution here is not that two dimensions differ from one, but which two dimensions, why, and what they imply for the direction of investment. On the AI side, the paper connects the measurement of task horizons

(Kwa et al., 2025; Ord, 2025), distillation (Hinton et al., 2015; Hsieh et al., 2023), and the economics of foundation-model competition (Korinek and Vipra, 2025; Acemoglu, 2025) into a single equilibrium framework; general-purpose-technology dynamics are as in Bresnahan and Trajtenberg (1995).

The rest of the paper proceeds as follows. Section 2 states the three aggregate facts. Section 3 develops the task technology and the representation theorem. Section 4 derives market payoffs. Section 5 derives the directional imitation technology. Section 6 characterizes the direction of frontier investment and the rotation result. Section 7 adds strategic competition between laboratories. Section 8 states the welfare wedge and the policy reversal. Section 9 confronts the model with aggregate evidence. Section 10 concludes. Proofs are in Appendix A; every formal claim in the paper is also verified line by line in a machine-checked companion script described in Appendix B.

2 Three Aggregate Facts

The theory is disciplined by three facts, each computable from public aggregate series. None requires micro data, and none is an identified causal estimate; their role, as in the quantitative-ideas literature (Bloom et al., 2020), is to pin the model's objects and to state in advance what would falsify it.

Fact 1 (The horizon frontier moves exponentially, and carries a reliability tax)**.** *The task duration at which frontier systems succeed 50% of the time has doubled roughly every seven months since 2019 (Kwa et al., 2025). Horizons measured at an 80% success requirement are roughly five times shorter than 50% horizons, with the same doubling time.*

The second half of Fact 1 is, to our knowledge, unexploited. A scalar model in which task success is governed by a constant per-minute hazard (Ord, 2025) predicts the 50%/80% horizon ratio *without free parameters*: if success on a task of length t is $e^{-t/T}$ for a model-specific T, then $t_{50}/t_{80} = \ln 0.5/\ln 0.8 \approx 3.1$, independent of T, and hence constant across models and time. An observed ratio near five is a rejection of the constant-hazard scalar benchmark in a specific direction: success decays with task length faster than geometrically at high reliability requirements. Section 3 shows this is exactly what task heterogeneity in *forgiveness* produces, and Section 9 uses the ratio to calibrate the coupling between the two dimensions.

Fact 2 (Imitation is fast, and directionally asymmetric)**.** *Open-weight models trail the frontier by roughly three to four months on aggregate capability indices. On autonomy time-horizon measures the lag is larger—mid-2025 open-weight systems matched the late-2024 frontier—and practitioner evidence consistently places the widest gaps in long-horizon agentic* reliability *rather than in demonstrated capability.*[1]

[1] Epoch AI, "Open-weight models lag state-of-the-art by around 3–4 months" (data insight, 2025–2026);

Fact 3 (Served tasks commoditize while frontier spending rises)**.** *The price of reaching a fixed benchmark score has fallen by one to two orders of magnitude per year (median estimate roughly* $50\times$ *per year across benchmarks), while frontier training expenditure has continued to grow.*[2]

Fact 3 is the puzzle. Facts 1 and 2 contain the resolution: the frontier is two-dimensional, and imitation erodes it asymmetrically. Any theory of the AI race should (i) generate rents that survive near-immediate imitation of demonstrated capability, (ii) predict which tasks commoditize and which do not, and (iii) say where frontier investment goes next. The model below is built to those specifications.

3 Task Technology and the Two-Dimensional Frontier

3.1 O-Ring Production with Retries

A task consists of n essential stages in series. Executing a stage is an *attempt*; an attempt by a system with reliability capability $q_F \in \mathbb{R}$ succeeds with probability $G_F(q_F - \varphi)$, where φ is the stage's local difficulty and G_F is a strictly increasing, differentiable CDF on $\mathbb{R}$. The task's *forgiveness* is the number of retries it permits: a stage may be attempted up to $1 + \ell$ times, $\ell \in \{0, 1, 2, \dots\}$, because errors are observable, reversible, or sandboxed enough to try again. The stage succeeds if any attempt does, so with $\pi = G_F(q_F - \varphi)$ the stage success probability is $S_\ell(\pi) = 1 - (1 - \pi)^{1+\ell}$, and the task's execution success probability is $S_\ell(\pi)^n$. Without retries this is Kremer's O-ring technology (Kremer, 1993); ℓ is the new margin, and it is the economic content of "forgiveness": version-controlled code allows many retries, a payment authorization or a lane change allows none.

A task also has a coordination requirement: the system must generate and hold together the n-step plan at all. We model this margin in reduced form: the plan succeeds with probability $G_H(q_H - h)$, where $h = \ln n$ is the task's *horizon*, q_H is the system's horizon capability, and G_H is a strictly increasing, differentiable CDF. Measuring horizon in logs matches how the frontier is measured empirically (Kwa et al., 2025) and makes Fact 1's exponential trend a linear drift in q_H. The task succeeds if the plan succeeds and every stage executes; a task is thus a tuple $\tau = (n, \varphi, \ell, \varepsilon)$, where ε is the failure probability the task's user tolerates.

METR, "Details about METR's preliminary evaluation of DeepSeek and Qwen models" (2025); Epoch AI, FrontierMath open-model analysis (lag of roughly seven months on the hardest tiers).

[2] Epoch AI, "LLM inference prices have fallen rapidly but unequally across tasks" (2025): declines of $9\times$–$900\times$ per year across benchmarks and thresholds.

3.2 The Reduced Fragility Requirement

Execution imposes a requirement on q_F that compounds with serial depth. Requiring $S_\ell(\pi)^n \geq 1-\varepsilon$ is equivalent to

$$q_F \;\geq\; f(\tau) \;\equiv\; \varphi + G_F^{-1}\Big(1 - \big\{1-(1-\varepsilon)^{1/n}\big\}^{1/(1+\ell)}\Big), \tag{1}$$

and we call $f(\tau)$ the task's *fragility*: the reliability capability the task demands. Fragility rises with serial depth n and local difficulty φ, falls with retries ℓ, and rises as the tolerated failure ε falls. Its structure is transparent in the logistic benchmark.

Lemma 1 (Log-linear fragility and the Gumbel kernel)**.** *Let G_F be the logistic CDF, $G_F(x) = 1/(1+e^{-x})$, and $h = \ln n$.*

(i) (Threshold.) *For $\varepsilon \leq 1/2$ and $n \geq 4^{1+\ell}\varepsilon$ (a task long relative to its forgiveness), the fragility requirement satisfies*

$$f(\tau) \;=\; \varphi + \frac{h + \ln(1/\varepsilon)}{1+\ell} + e(\tau), \qquad |e(\tau)| \;\leq\; 2\varepsilon + 3\Big(\frac{\varepsilon}{n}\Big)^{1/(1+\ell)}.$$

The reliability requirement is linear in log-horizon with slope $1/(1+\ell)$: serial depth taxes reliability, and forgiveness divides the tax.

(ii) (Kernel.) *Fix $u \in \mathbb{R}$ and let $q_F = \varphi + (\ln n + u)/(1+\ell)$. Then as $n \to \infty$,*

$$S_\ell\big(G_F(q_F - \varphi)\big)^n \;\longrightarrow\; \exp\big(-e^{-u}\big),$$

with error $O\big(n^{-1/(1+\ell)}\big)$. Equivalently, execution success at capability q_F converges to the Gumbel kernel $K_\ell(q_F - f) = \exp\{-e^{-(1+\ell)(q_F - f)}\}$ with location equal to the fragility f of the task at ε-scale and scale $1/(1+\ell)$.

Lemma 1 does two jobs. First, it replaces an assumption with a derivation: smooth "success curves" $P(q-f)$ of the kind task-based models posit are the large-n shape of O-ring-with-retries production, and their steepness is not free— it is $1+\ell$, the task's forgiveness. Forgiving tasks flip from infeasible to routine quickly; unforgiving tasks have long, shallow reliability ramps. Second, it reduces a four-parameter task τ to a two-dimensional requirement $(h, f) \in \mathbb{R}^2$: horizon, and the reliability demanded at that horizon. These two coordinates are the task frontier's space.

3.3 When Is a Scalar Ladder Enough?

The reduced requirements answer the first-order modeling question—when is "model quality" one number?—as a theorem rather than a taste. Say a system *serves* a task if it meets both

requirements: $q_H \geq h$ and $q_F \geq f$.[3] A *scalar representation* of a task economy $\mathcal{T} = \{(h_t, f_t)\}$ is a pair of indices $\iota : \mathcal{T} \to \mathbb{R}$ and $\tilde{\iota} : \mathbb{R}^2 \to \mathbb{R}$ such that any system q serves task t if and only if $\tilde{\iota}(q) \geq \iota(t)$.

Theorem 1 (Representation)**.** *A task economy $\mathcal{T}$ admits a scalar representation if and only if its requirement pairs are totally ordered by the componentwise order (a chain): there is no pair of tasks with $h_1 < h_2$ and $f_1 > f_2$. In particular:*

(i) If all tasks share the same forgiveness and local difficulty (ℓ, φ), then f is a strictly increasing function of h along (1), $\mathcal{T}$ is a chain, and the scalar quality ladder is exact.

(ii) If forgiveness is heterogeneous, the economy generically contains unordered pairs—a long forgiving task and a short unforgiving one—and no scalar index can rank systems: there exist systems q, q' neither of which serves a superset of the other's tasks.

The scalar ladder is thus the special case of *homogeneous forgiveness*. It fails on precisely the comparisons that organize the AI market: multi-hour coding with tests and version control (large h, large ℓ) against payment authorization, medication dosing, or a driving maneuver (small h, $\ell = 0$). Once such pairs carry value, "better" is not a number, and the direction of progress is an economic choice. The rest of the paper is about who chooses it and whether they choose well.

3.4 Task Value and Boundary Statistics

Let $a(h, f) \geq 0$ be the value density over requirements: the flow surplus from tasks at (h, f) if they are served. A system $q = (q_H, q_F)$ generates gross surplus

$$W(q) \;=\; \iint a(h,f)\, P_H(q_H - h)\, P_F(q_F - f)\, dh\, df, \tag{2}$$

where P_H, P_F are strictly increasing success kernels; by Lemma 1, $P_F = K_\ell$ is the derived Gumbel shape, and we carry general kernels for robustness. Write W_R for the integral restricted to a region R. The marginal values of capability are the *boundary values*

$$B_H(q) = \frac{\partial W}{\partial q_H} = \iint a\, g_H\, P_F, \qquad B_F(q) = \frac{\partial W}{\partial q_F} = \iint a\, P_H\, g_F, \tag{3}$$

with g_i the kernel densities, and the cross statistic

$$C(q) \;=\; \frac{\partial B_H}{\partial q_F} \;=\; \frac{\partial B_F}{\partial q_H} \;=\; \iint a\, g_H\, g_F \;\geq\; 0. \tag{4}$$

[3] The deterministic-threshold statement is for transparency; by Lemma 1(ii) it is the sharp-kernel limit of the probabilistic model, and the theorem's logic—nestedness of served sets—is unchanged under strictly monotone kernels.

B_H prices the value mass just beyond the horizon boundary, B_F the mass just beyond the reliability boundary, and C—always nonnegative—the mass near the corner where both bind. For a finite capability step $\Delta \geq 0$,

$$W(q+\Delta) - W(q) = \int_0^{\Delta_H} B_H(q_H + u,\, q_F)\, du + \int_0^{\Delta_F} B_F(q_H + \Delta_H,\, q_F + v)\, dv. \tag{5}$$

Equation (4) is worth pausing on: it says that pushing the horizon boundary *mechanically raises* the marginal value of reliability whenever value sits near both boundaries. Long tasks, once reachable in horizon, wait on reliability. This derived complementarity—not an assumption—is what will drive the rotation of investment in Section 6.

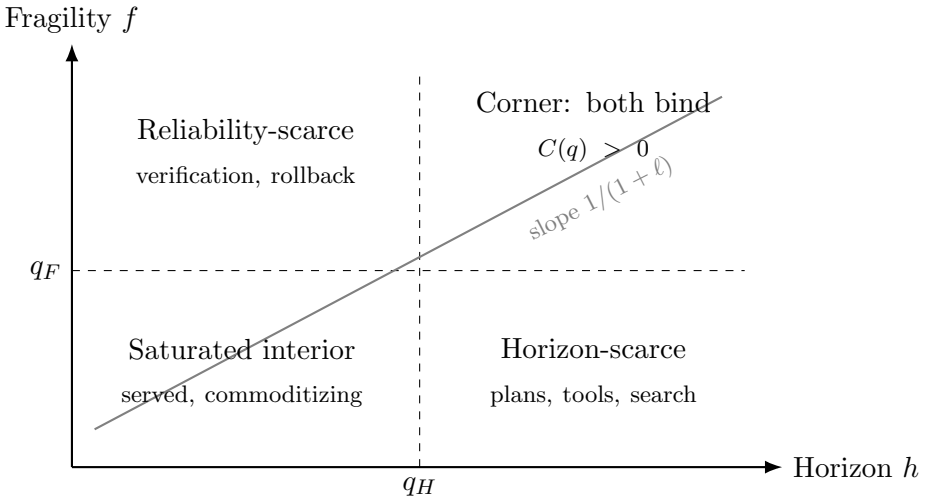

Figure 1: The task frontier. Tasks are points (h, f); a family with common forgiveness ℓ lies on a line of slope $1/(1+\ell)$ (Lemma 1). Capability $q = (q_H, q_F)$ serves the lower-left region; boundary values B_H, B_F price the two frontier directions.

4 Market Equilibrium: Profits Are Gaps

Task-based theories of AI competition typically posit profit functions—a rent from being ahead, an outside option that erodes with "competitive pressure." Here both are equilibrium objects, and their structure does real work later, so we derive them.

Each task (h, f) is a market. The buyer values successful completion at $a(h, f)$ per period, so system k delivers expected gross value $v_k(h, f) = a(h, f)\, P_H(q_H^k - h) P_F(q_F^k - f)$ at serving cost $c_k(h, f)$; let $s_k = v_k - c_k$ denote delivered net surplus. Sellers set task-specific prices; buyers choose sellers. This is asymmetric Bertrand competition with vertically differentiated suppliers.

Lemma 2 (Bertrand in tasks)**.** *In every task's pricing equilibrium:*

(i) (Allocation.) *The task is served by the firm with the highest delivered net surplus $s_k(h, f)$. A lower-cost follower m displaces the frontier firm L exactly where its cost advantage exceeds the value gap: $c_L - c_m \geq v_L - v_m$.*

(ii) (Pricing.) *The winner's margin on the task is the surplus gap over the runner-up, $s_{(1)}(h, f) - s_{(2)}(h, f)$.*

(iii) (Profit is a gap integral.) *If the frontier firm has capability $q^L \geq q^m$ componentwise, equal serving costs, and the best alternative on every task is the fringe system q^m, its flow profit is*

$$\Pi\big(q^L, q^m\big) \;=\; W(q^L) - W(q^m) \;=\; \int_0^{\Delta_H} B_H(q_H^m + u,\, q_F^m)\, du + \int_0^{\Delta_F} B_F(q_H^L,\, q_F^m + v)\, dv,$$

where $\Delta = q^L - q^m$: boundary values integrated over the capability gap.

(iv) (Saturation.) *As the fringe closes the gap in a region ($q^m \to q^L$ there), the frontier firm's margins in that region converge to zero while delivered surplus does not: the region is saturated—served, valuable, and rentless.*

Three consequences organize what follows. First, part (i) is the sorting rule that, in earlier task-frontier drafts, appeared as a standalone observation; here it is the equilibrium allocation, and "cheap models win the saturated interior" is a location statement about where value gaps are small. Second, part (iii) says *rents live in the gap*: the frontier firm's profit is not the value of what it serves but the boundary-value mass between itself and its best imitator. Current revenue and frontier rent are different objects—a region can be economically huge and competitively worthless. Third, the envelope property of (iii) gives the frontier firm's marginal value of capability: $\partial\Pi/\partial q_i^L = B_i(q^L)$, the boundary value at its own frontier, *so long as the gap persists.* What the fringe does to the gap over time is therefore the entire strategic question, and it is where we go next.

5 Imitation Technology: The Tail Cannot Be Distilled

The fringe in Lemma 2 was a capability q^m; we now give it a technology. Followers learn from what the frontier deploys: successful trajectories, tool scaffolds, reasoning traces, evaluation results—the raw material of distillation (Hinton et al., 2015; Hsieh et al., 2023). The question is what observation can and cannot teach. The answer is asymmetric in exactly the two dimensions of the task frontier.

Horizon is a blueprint good. The plan structure that completes a long task—the decomposition, the tool calls, the checkpoints—is embodied in any successful trajectory. One demonstration reveals it, in Arrow's sense that the information is disclosed by the product itself (Arrow,

1962). We model this as: access to any positive flow of demonstrations moves q_H^m toward q_H^L at rate at least $\underline{\lambda}_H > 0$, a constant of access, not of statistics.

Reliability is a tail statistic. A follower cannot serve an unforgiving task by exhibiting a plan; it must *be* reliable at the task's tolerated failure rate ε, and—because unforgiving tasks are precisely those where failures are costly—its system must be certified at that rate before deployment, by itself, its customers, or a regulator. Certification is a hypothesis-testing problem, and it has a sample-complexity floor.

Theorem 2 (The tail cannot be distilled)**.** *(i)* (Sample floor.) *Any certification procedure that, with error probabilities at most $\delta < 1/2$, distinguishes a system with failure rate $\leq \varepsilon/2$ from one with failure rate $\geq \varepsilon$ requires*

$$N(\varepsilon) \;\geq\; \frac{2(1-2\delta)^2}{\mathrm{KL}\big(\mathrm{Bern}(\varepsilon/2)\,\|\,\mathrm{Bern}(\varepsilon)\big)} \;\geq\; \frac{4(1-2\delta)^2}{\varepsilon}$$

independent task-level observations for $\varepsilon \leq 1/2$, since $\mathrm{KL} \leq \chi^2 = \varepsilon/\{4(1-\varepsilon)\} \leq \varepsilon/2$ there; the sharp small-ε coefficient is $4/(1-\ln 2) \approx 13$ times $(1-2\delta)^2$.

(ii) (Exponential wall.) *At the fragility frontier, the tolerated failure rate of the marginal task is $\varepsilon(q_F) \asymp n\, e^{-(1+\ell)(q_F-\varphi)}$ (Lemma 1). If demonstrations of frontier behavior on such tasks arrive at a bounded flow D per period, the time for a follower to certify at the frontier's reliability level satisfies*

$$T_{\mathrm{imitate}}(q_F) \;\geq\; \frac{N(\varepsilon(q_F))}{D} \;\gtrsim\; \frac{e^{(1+\ell)(q_F-\varphi)}}{D\,n},$$

growing exponentially in the position of the reliability frontier. The implied gap-decay rate $\lambda_F(q_F)$ falls toward zero as the frontier advances, while $\lambda_H \geq \underline{\lambda}_H > 0$.

(iii) (Directional appropriability.) *Consequently, for any discount rate $r > 0$, the fraction of a capability lead's value that survives imitation, $\lambda \mapsto r/(r+\lambda)$ per Section 6, is strictly larger in the reliability direction than in the horizon direction, and the asymmetry widens as the reliability frontier advances.*

The one-line version: *a plan can be copied from one observation; a failure rate must be estimated from many.* Appropriability here is not a patent, a secret, or a complementary asset (Teece, 1986; Anton and Yao, 1994)—it is a statistical property of the task being served. Forgiving, observable, high-volume tasks teach imitators everything and protect nothing. Unforgiving tasks are self-protecting: the same scarcity of tolerated failures that makes them hard makes them slow to copy. Two remarks bound the claim honestly. First, the floor in (i) binds *certification from behavior*; a follower could instead re-derive reliability from first principles—the theorem prices imitation, not independent invention, which is what the investment technology in the next section is. Second, public verification infrastructure—shared

incident registries, standardized high-reliability evaluations, certification authorities—directly attacks the supply side of the bound by raising the usable observation flow D: it is, in this precise sense, imitation policy for the reliability dimension. Section 8 returns to this.

6 The Direction of Frontier Investment

We now put the three derived objects—gap profits, directional imitation, and the cross-boundary complementarity—into motion. To characterize direction cleanly we study a single frontier laboratory against the imitative fringe; strategic interaction between laboratories is Section 7.

6.1 The Leader–Fringe Problem

Time is continuous, the discount rate is $r > 0$. The state is the capability gap $\Delta = (\Delta_H, \Delta_F) = q^L - q^m \geq 0$. The laboratory's investment $x_i \geq 0$ raises its own capability in direction i; the fringe closes each gap at its imitation rate from Theorem 2:

$$\dot{\Delta}_i \;=\; x_i - \lambda_i \Delta_i, \qquad i \in \{H, F\}, \qquad \lambda_H > \lambda_F \geq 0. \tag{6}$$

By Lemma 2(iii), flow profit is the gap integral of boundary values. We work in the *local regime* in which the boundary values are constant over the width of the gap,[4] so $\Pi(\Delta) = B_H \Delta_H + B_F \Delta_F$, with $B_i = B_i(q^L)$ evolving slowly as the frontier itself moves. Investment costs are $c(x_H) + c(x_F)$, increasing and convex.

Theorem 3 (Shadow values and rotation)**.** *(i)* (Shadow values.) *The value function is linear:*

$$V(\Delta) = \mu_H \Delta_H + \mu_F \Delta_F + v_0, \qquad \mu_i \;=\; \frac{B_i}{r + \lambda_i}:$$

a unit of capability lead in direction i is worth its boundary value, discounted by interest plus the rate at which imitation erases it. *Optimal investment equates marginal cost to μ_i (interior case $x_i^* = c'^{-1}(\mu_i)$; with a linear budget, all investment goes to the direction with the higher μ_i/c_i).*

(ii) (Rotation.) *Suppose the frontier starts horizon-directed, $\mu_H/c_H > \mu_F/c_F$, and along the induced horizon push the scarcity ratio rises, $\frac{d}{dq_H}\big(B_F/B_H\big) > 0$—which holds whenever $C(q) > 0$ and $\partial B_H/\partial q_H \leq 0$, i.e., whenever unlocked long tasks wait on reliability and pure-horizon value mass depletes—and eventually exceeds $\rho^* \equiv \dfrac{(r + \lambda_F)\, c_F}{(r + \lambda_H)\, c_H}$. Then there*

[4] Formally, an approximation whose error is second order in the gap: by (5), $|\Pi(\Delta) - B \cdot \Delta|$ is bounded by the variation of the boundary values over the gap rectangle, i.e., by the curvature of W (including the cross statistic C) times $\|\Delta\|^2/2$. The regime is the empirically relevant one: measured frontier–fringe gaps (Fact 2) are months wide, while the value distribution moves on the scale of years.

is a finite rotation time T^*, before which investment is horizon-directed and after which it is reliability-directed.

(iii) (Imitation steers.) *The rotation threshold ρ^* is strictly decreasing in λ_H and increasing in λ_F: faster imitation of the horizon dimension brings the rotation* earlier*, and faster diffusion of reliability postpones it. Under single crossing of B_F/B_H, T^* inherits these comparative statics.*

Part (i) is the paper's price theory of direction: the market prices frontier directions by *persistence-adjusted* boundary values $B_i/(r+\lambda_i)$, not by boundary values. Part (ii) turns the derived complementarity (4) into dynamics. A horizon push does two things at once: it depletes the stock of tasks whose binding constraint is horizon, and it unlocks—in the horizon dimension—long tasks that now wait on reliability. Both forces raise B_F/B_H, so the direction of investment must eventually rotate toward the unforgiving boundary. Part (iii) is the result we regard as the model's signature. In scalar models, imitation is a tax on innovation: it lowers the return and can only slow the race (or, in escape-competition settings, speed it along the same axis). Here imitation has a *steering* effect: commoditization of demonstrated capability is precisely what pushes the frontier toward reliability. The fringe does not slow the leader down so much as tell it where to go.

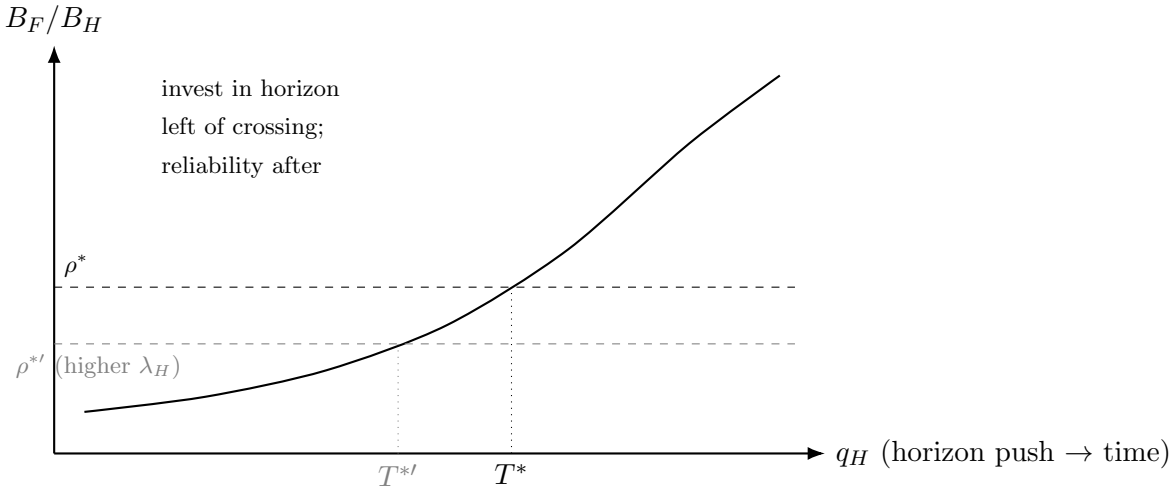

Figure 2: Rotation. Along a horizon push the scarcity ratio B_F/B_H rises (depletion plus the derived complementarity $C > 0$); investment rotates to the reliability boundary when it crosses the persistence-adjusted threshold ρ^*. Faster horizon imitation (higher λ_H) lowers the threshold and brings the rotation earlier.

The rotation result reads directly onto the industry's trajectory: a phase in which frontier competition is about demonstrated capability—longer benchmarks, agents, scaffolds—followed, as those capabilities commoditize (Facts 2 and 3), by a phase in which it is about reliability

in unforgiving deployments: verification, monitoring, rollback, certified autonomy. Section 9 asks where on this diagram the observable aggregates place us.

7 Strategic Competition: Racing on the Same Boundary

A scalar race leaves rivals only one question: how hard to run. A two-dimensional frontier adds a question a scalar model cannot pose: *where* to run—and whether competitors split the frontier or crowd onto the same boundary.

Let two frontier laboratories simultaneously commit investment programs to a direction $d_i \in \{H, F\}$. Racing alone on boundary i is worth the leader–fringe value of that direction, V_i (by Theorem 3, V_i is increasing in $\mu_i = B_i/(r+\lambda_i)$); racing head-to-head on the same boundary splits and dissipates the prize to σV_i each, $\sigma \in (0, \frac{1}{2}]$. Order the directions so $V_F \geq V_H$—by Theorem 2 this is the configuration imitation asymmetry produces even when raw boundary values are comparable.

Theorem 4 (Herding on the moat)**.** *In the direction game:*

(i) If $\sigma V_F \geq V_H$, racing on the reliability boundary is a dominant strategy: the unique equilibrium is herding*, both laboratories on the moat dimension, each accepting the split prize σV_F.*

(ii) If $\sigma V_F < V_H \leq V_F$, the pure equilibria are the two differentiation *profiles (one laboratory per boundary), and there is a mixed equilibrium in which each races on F with probability $p^* = \dfrac{V_F - \sigma V_H}{(V_F - \sigma V_H) + (V_H - \sigma V_F)}$, strictly increasing in V_F/V_H.*

(iii) Because V_F/V_H is increasing in λ_H and decreasing in λ_F, faster imitation of demonstrated capability expands the herding region: sufficiently asymmetric imitation makes both *rivals race on the same unforgiving frontier despite splitting its rents.*

The result inverts the differentiation intuition. In classic quality competition, rivals spread out to soften price competition (Shaked and Sutton, 1982); a naive port to innovation says rival laboratories should divide the frontier between them. Theorem 4 says the moat is worth sharing: when the horizon dimension is copied in months, half of a defensible reliability lead beats all of an indefensible capability lead. The prediction is concentration of frontier R&D portfolios on the same reliability-heavy direction—simultaneous pivots by rival laboratories toward agentic reliability, verification, and high-stakes deployment—rather than specialization, exactly when open-weight catch-up is fastest.

8 Welfare: The Direction Wedge and a Policy Reversal

8.1 Whose Direction Is Right?

The planner and the market price the same two directions differently, and the difference is the paper's welfare economics. Society keeps a capability advance forever: once the frontier reaches q, surplus $W(q)$ is available regardless of who serves the tasks, and the fringe's subsequent catch-up is diffusion—a transfer from margins to buyers—not destruction. The planner's marginal value of a unit of direction-i capability is therefore B_i/r. The laboratory, by Theorem 3, values it at $B_i/(r+\lambda_i)$: it only earns the gap, and the gap decays.

Theorem 5 (The direction wedge is the imitation asymmetry)**.** *(i) The market's valuation of direction i relative to the planner's is $r/(r+\lambda_i)$, and the* relative direction distortion *is*

$$\frac{\mu_H/\mu_F}{(B_H/r)/(B_F/r)} = \frac{r+\lambda_F}{r+\lambda_H} < 1 \quad \Longleftrightarrow \quad \lambda_H > \lambda_F.$$

The market undervalues exactly the dimension that is easy to imitate. Its sign and magnitude are pinned by an observable: the asymmetry of imitation lags (Fact 2).

(ii) Measured by surplus W, the market therefore over-*rotates toward the reliability boundary relative to the planner—it is drawn by persistence, not value. If, in addition, failures on unforgiving tasks carry uninternalized social harm, the planner's reliability boundary value scales to $\theta_s B_F$ with $\theta_s > 1$, and the two distortions oppose: the net direction bias has the sign of $\theta_s - \dfrac{r+\lambda_H}{r+\lambda_F}$. Whether the market is too safety-directed or not safety-directed enough is not an ideological constant; it is a horse race between an externality and an imitation asymmetry, both in principle measurable.*

Part (ii) deserves emphasis because it cuts against both popular narratives at once. The pure appropriability logic says markets *favor* reliability—it is the moat—so laissez-faire innovation drifts toward the unforgiving frontier faster than surplus alone justifies. The pure externality logic says markets underinvest in safety. Both forces are real in the model, they load on the same margin with opposite signs, and the theory's contribution is to say precisely which observables settle the race.

8.2 Two Safety Policies, Opposite Market Structures

Take the quadratic-cost leader–fringe economy of Theorem 3 ($x_i^* = \mu_i$, steady-state gap $\Delta_i^* = x_i^*/\lambda_i$), and compare two instruments aimed at "more reliable AI."

Liability. A liability rule or regulatory standard on unforgiving tasks makes sellers internalize failures, scaling the private reliability boundary value to θB_F, $\theta > 1$.

Public verification infrastructure. Shared high-reliability evaluation suites, incident registries, and certification institutions raise the observation flow that Theorem 2 shows is the binding constraint on imitating reliability: they raise λ_F.

Corollary 1 (Policy reversal)**.** *In steady state, with frontier-rent concentration measured by the reliability component of the gap profit,* $\Pi_F^* = B_F \Delta_F^* = B_F^2/\{(r+\lambda_F)\lambda_F\}$*:*

(i) Liability raises reliability investment ($\partial x_F^*/\partial\theta > 0$) *and* thickens *the moat:* $\partial\Delta_F^*/\partial\theta > 0$*, concentration rises.*

(ii) Verification infrastructure thins *the moat and diffuses reliability:* $\partial\Delta_F^*/\partial\lambda_F < 0$*, so the fringe's reliability rises toward the frontier. The cost is the private frontier incentive itself,* $\partial x_F^*/\partial\lambda_F < 0$*.*

Both policies deliver "safer AI"; they move market structure in opposite directions.

The corollary gives competition and safety authorities a common language. Liability is safety policy that doubles as concentration policy: it pays the frontier firm in exactly the currency—persistent, hard-to-imitate reliability—that Theorem 2 says cannot be competed away. Verification infrastructure is competition policy for the reliability dimension: it attacks the sample-complexity wall directly, converting the moat into a commons. The model does not rank them—the ranking depends on θ_s, r, and the λ's—but it insists they are not the same policy, and that treating "AI safety investment" as a single quantity conflates a frontier incentive with a diffusion externality.

9 Quantitative Discipline

The model's load-bearing objects are few: the coupling slope $1/(1+\ell)$, the boundary values B_i, and the imitation rates λ_i. Each is disciplined by one of the aggregate facts of Section 2. We emphasize what this section is and is not: it is measurement that constrains a theory, in the tradition of Bloom et al. (2020); it identifies no causal parameter, and the theory carries the causal weight.

9.1 The 50%/80% Ratio as a Specification Test

The scalar constant-hazard benchmark makes a parameter-free prediction. If task success is $e^{-t/T}$ in task length t (Ord, 2025)—the continuous O-ring limit with homogeneous stages and no retry heterogeneity—then measured horizons at reliability requirements 50% and 80% satisfy

$$\frac{t_{50}}{t_{80}} = \frac{\ln 0.5}{\ln 0.8} \approx 3.11,$$

for every model, at every date, independent of the hazard T; the same invariance holds for any common retry depth (the ratio is unchanged by ℓ). The measured ratio is roughly *five* (Kwa et al., 2025). The excess is a rejection of the scalar benchmark in the model's direction: at high reliability requirements, feasible horizons contract faster than a homogeneous-hazard economy allows, exactly what a task mix heterogeneous in forgiveness produces—raising the required reliability disproportionately excludes the unforgiving tasks. The 50/80 ratio is thus a one-number specification test separating scalar from two-dimensional task economies, and its evolution as the reliability frontier advances is a standing falsification opportunity for the model.

9.2 The Imitation Asymmetry and the Size of the Wedge

Fact 2 gives gap half-lives by dimension: capability indices are reproduced in $\tau_H \approx 3$–4 months; documented autonomy time-horizon lags run six to twelve months, and high-reliability agentic performance sits beyond them, so we take $\tau_F \approx 14$ months as an illustrative value for the high-reliability margin. Reading $\lambda_i = \ln 2/\tau_i$ and a monthly discount rate of one percent, the relative direction distortion of Theorem 5 is

$$\frac{r+\lambda_F}{r+\lambda_H} \;=\; \frac{0.01+\ln 2/14}{0.01+\ln 2/4} \;\approx\; 0.3:$$

per unit of boundary value, the market prices a horizon lead at roughly one-third of a reliability lead. Equivalently, at equal marginal costs the rotation threshold ρ^* of Theorem 3 sits near one-third: frontier investment tips toward the unforgiving boundary as soon as reliability-boundary value per dollar reaches a third of horizon-boundary value per dollar. An asymmetry this large makes the model's herding region (Theorem 4) wide: rival laboratories should be observed pivoting to the *same* reliability-heavy direction, not specializing—consistent with the simultaneous industry turn toward agentic reliability, verification, and certified deployment as capability benchmarks commoditized.

9.3 What Would Falsify the Theory

The model stakes out four aggregate predictions. (1) *Widening asymmetry*: as the reliability frontier advances, the open-weight lag on high-reliability measures should grow relative to the capability-index lag (the exponential wall of Theorem 2); convergence of the two lags would falsify the mechanism. (2) *Rotation*: frontier laboratories' investment mix should shift toward reliability-directed projects following accelerations in capability commoditization, with the timing ordered by Fact 3's price-collapse episodes. (3) *Herding, not specialization*: rival laboratories' frontier portfolios should become more similar as imitation asymmetry widens. (4) *Policy cross-check*: where public verification infrastructure expands (standardized high-stakes evaluations, incident registries), fringe entry into fragile-task markets should follow and

frontier margins there should compress. Each prediction uses only public aggregates; none has, to our knowledge, been assembled, and any of them could kill the model.

10 Conclusion

The economically relevant state of AI is a frontier in a task space with two derived dimensions: how long a chain of actions a task requires, and how unforgiving it is of error along the way. One O-ring-with-retries primitive delivers the geometry: serial depth taxes reliability, forgiveness divides the tax, and the scalar quality ladder is the knife-edge of homogeneous forgiveness. On this frontier, competition prices capability leads by their persistence, and persistence is statistics: plans are copied from single demonstrations, failure rates only from many. So rents are competed away along the horizon dimension and defended along the reliability dimension; investment rotates from the former to the latter as followers close in; rivals herd onto the moat; and the direction of AI innovation is distorted by an observable—the asymmetry of imitation lags—in a horse race with the safety externality that runs the other way. Cheap models and rising frontier spending are not a paradox. They are the same force seen from two sides: imitation erases the value of what can be demonstrated, and thereby concentrates the race on what cannot.

A Proofs

A.1 Proof of Lemma 1

Part (i). The requirement $S_\ell(\pi)^n \geq 1-\varepsilon$ with $S_\ell(\pi) = 1-(1-\pi)^{1+\ell}$ is equivalent to $1-(1-\pi)^{1+\ell} \geq (1-\varepsilon)^{1/n}$, i.e. to $\pi \geq 1-g^{1/(1+\ell)}$ with $g \equiv 1-(1-\varepsilon)^{1/n}$; since G_F is strictly increasing this is (1). For the error bound, write $y = g^{1/(1+\ell)}$ and, with logistic G_F,

$$f(\tau)-\varphi = \operatorname{logit}(1-y) = -\ln y + \ln(1-y) = \frac{1}{1+\ell}\ln\frac{1}{g} + \ln(1-y),$$

so

$$e(\tau) = \underbrace{\frac{1}{1+\ell}\ln\frac{\varepsilon/n}{g}}_{(a)} + \underbrace{\ln(1-y)}_{(b)}.$$

The integral representation $g = \int_0^\varepsilon \frac{1}{n}(1-t)^{\frac{1}{n}-1}dt$ with integrand between $1/n$ and $\frac{1}{n}(1-\varepsilon)^{\frac{1}{n}-1} \leq \frac{1}{n(1-\varepsilon)}$ gives the sandwich

$$\frac{\varepsilon}{n} \;\leq\; g \;\leq\; \frac{\varepsilon}{n(1-\varepsilon)}. \tag{7}$$

Hence $(a) \in [-\frac{1}{1+\ell}\ln\frac{1}{1-\varepsilon}, 0]$, and for $\varepsilon \leq 1/2$, $|(a)| \leq \ln\frac{1}{1-\varepsilon} \leq \frac{\varepsilon}{1-\varepsilon} \leq 2\varepsilon$. For (b): the condition $n \geq 4^{1+\ell}\varepsilon$ means $(\varepsilon/n)^{1/(1+\ell)} \leq 1/4$, and (7) with $\varepsilon \leq 1/2$ gives $y \leq \left(\frac{2\varepsilon}{n}\right)^{1/(1+\ell)} =$

$2^{1/(1+\ell)}\left(\frac{\varepsilon}{n}\right)^{1/(1+\ell)} \le 2\cdot\frac{1}{4} = \frac{1}{2}$. On $y \le 1/2$, $|\ln(1-y)| \le (2\ln 2)\, y \le 1.39\cdot 2^{1/(1+\ell)}(\varepsilon/n)^{1/(1+\ell)} \le 3(\varepsilon/n)^{1/(1+\ell)}$. Adding the two bounds gives the claim. Monotonicities: g is increasing in ε and decreasing in n; y is increasing in g and decreasing in ℓ; f is decreasing in y-threshold direction as displayed; and φ enters additively. □

Part (ii). Let $x_n = q_F - \varphi = (\ln n + u)/(1+\ell)$ and $y_n = (1-\pi_n)^{1+\ell}$ with $\pi_n = G_F(x_n)$. For the logistic, $1-\pi_n = e^{-x_n}/(1+e^{-x_n})$, so

$$n\, y_n = n\, e^{-(1+\ell)x_n}\left(1+e^{-x_n}\right)^{-(1+\ell)} = e^{-u}\left(1+e^{-x_n}\right)^{-(1+\ell)}, \qquad e^{-x_n} = \left(\frac{e^{-u}}{n}\right)^{1/(1+\ell)} \to 0.$$

Thus $ny_n = e^{-u}\{1 + O(n^{-1/(1+\ell)})\}$. Then $S_\ell(\pi_n)^n = (1-y_n)^n = \exp\{n\ln(1-y_n)\} = \exp\{-ny_n + O(ny_n^2)\}$, and $ny_n^2 = (ny_n)\, y_n = O(1/n)$. Combining, $S_\ell(\pi_n)^n = \exp(-e^{-u})\{1 + O(n^{-1/(1+\ell)})\}$. The kernel form follows by setting $u = (1+\ell)(q_F - f)$ with $f = \varphi + (\ln n)/(1+\ell)$. □

A.2 Proof of Theorem 1

Throughout, "serves" means $q_H \ge h_t$ and $q_F \ge f_t$; take $\mathcal{T}$ finite (or compact in (h,f)).

(Chain ⇒ scalar.) Suppose the requirement pairs form a chain. Define $\iota(t) = h_t + f_t$; along a chain, $t \prec t'$ (componentwise, one strict) implies $\iota(t) < \iota(t')$, so ι is a strict order embedding. The set of tasks a system q serves is a *lower set* of the chain: if q serves t' and $t \preceq t'$ then q serves t. Hence the served set is an initial segment, and defining $\tilde{\iota}(q) = \max\{\iota(t) : q \text{ serves } t\}$ ($-\infty$ if none; the max exists by finiteness/compactness), q serves t iff $\iota(t) \le \tilde{\iota}(q)$.

(Unordered pair ⇒ no scalar.) Let $h_1 < h_2$ and $f_1 > f_2$. System $A = (h_2, f_2)$ serves t_2 but not t_1 (it fails $f_1 > f_2$); system $B = (h_1, f_1)$ serves t_1 but not t_2 (it fails $h_2 > h_1$). A scalar representation would require $\iota(t_2) \le \tilde{\iota}(A) < \iota(t_1)$ and $\iota(t_1) \le \tilde{\iota}(B) < \iota(t_2)$, a contradiction.

(i) With common $(\ell, \varphi, \varepsilon)$, (1) makes f a strictly increasing function of n, hence of $h = \ln n$: the requirement set lies on an increasing curve and is a chain. *(ii)* With two forgiveness classes $\ell_1 > \ell_2$, Lemma 1(i) puts their requirement pairs on lines of slopes $1/(1+\ell_1) < 1/(1+\ell_2)$; whenever the economy contains a long task of the forgiving class and a short task of the unforgiving class with f-ranking reversed—e.g. multi-hour tested code versus a one-shot payment authorization—the pair is unordered and the previous paragraph applies. □

A.3 Proof of Lemma 2

Fix a task and drop its index. Sellers simultaneously post prices; the buyer accepts the offer maximizing $v_k - p_k$, so seller k wins with any price $p_k \le v_k - (v_j - p_j)$ against best rival offer j. Pricing below cost is weakly dominated; in equilibrium the runner-up offers $p_{(2)} = c_{(2)}$ (any higher loses for sure against the standard undercutting argument, any lower is dominated), and the winner posts the highest winning price $p_{(1)} = v_{(1)} - (v_{(2)} - c_{(2)})$, earning margin

$s_{(1)} - s_{(2)} \geq 0$. This proves (ii), and (i) since winning requires $s_k \geq s_j$ for all j; the displacement condition is $s_m \geq s_L$ rearranged. For (iii), with equal serving costs the margin on each task is $v_L - v_m = a(h,f)\{P_H(q_H^L - h)P_F(q_F^L - f) - P_H(q_H^m - h)P_F(q_F^m - f)\} \geq 0$ (componentwise dominance and monotone kernels); integrating over tasks gives $\Pi = W(q^L) - W(q^m)$, and (5) is the fundamental theorem of calculus applied along the horizontal-then-vertical path, using (3). Part (iv) is continuity of W. □

A.4 Proof of Theorem 2

(i) Let $P = \mathrm{Bern}(\varepsilon/2)^{\otimes N}$, $Q = \mathrm{Bern}(\varepsilon)^{\otimes N}$. Any test ψ satisfies $\Pr_P(\psi = 1) + \Pr_Q(\psi = 0) \geq 1 - \mathrm{TV}(P,Q)$, and by Pinsker $\mathrm{TV}(P,Q) \leq \sqrt{\mathrm{KL}(P\|Q)/2} = \sqrt{N\,\mathrm{KL}/2}$ with $\mathrm{KL} = \mathrm{KL}(\mathrm{Bern}(\varepsilon/2)\|\mathrm{Bern}(\varepsilon))$. Error probabilities at most δ therefore require $2\delta \geq 1 - \sqrt{N\,\mathrm{KL}/2}$, i.e. $N \geq 2(1-2\delta)^2/\mathrm{KL}$. The χ^2 bound $\mathrm{KL}(P_1\|Q_1) \leq \chi^2(P_1\|Q_1)$ gives, for Bernoulli,

$$\chi^2 = \frac{(\varepsilon/2)^2}{\varepsilon} + \frac{(\varepsilon/2)^2}{1-\varepsilon} = \frac{\varepsilon}{4(1-\varepsilon)} \;\leq\; \frac{\varepsilon}{2} \quad (\varepsilon \leq 1/2),$$

hence $N \geq 4(1-2\delta)^2/\varepsilon$. As $\varepsilon \to 0$, $\mathrm{KL}/\varepsilon \to (1 - \ln 2)/2$, giving the sharp coefficient $4/(1 - \ln 2) \approx 13.0$.

(ii) By Lemma 1(ii), the failure rate of the frontier system on a task at its fragility frontier is $1 - \exp\{-e^{-(1+\ell)(q_F - f)}\} \asymp n\, e^{-(1+\ell)(q_F - \varphi)}$ at the marginal served task, so the tolerated ε a follower must certify is of this order. With at most D usable independent observations per period, $T_{\text{imitate}} \geq N(\varepsilon)/D \geq 4(1-2\delta)^2 e^{(1+\ell)(q_F-\varphi)}/(Dn)$ up to the constant absorbed in $\asymp$. The implied exponential-decay rate of the reliability gap is $\lambda_F(q_F) \propto 1/T_{\text{imitate}} \to 0$. The horizon dimension needs one successful trajectory per plan structure, so its lag is bounded by access delay: $\lambda_H \geq \underline{\lambda}_H > 0$.

(iii) Immediate from (ii) and the definition of the persistence factor. □

A.5 Proof of Theorem 3

(i) Consider $V(\Delta) = \mu_H \Delta_H + \mu_F \Delta_F + v_0$. The HJB equation is

$$rV(\Delta) = \max_{x \geq 0} \Big\{ B_H \Delta_H + B_F \Delta_F - c(x_H) - c(x_F) + \sum_i V_i'(\Delta)\,(x_i - \lambda_i \Delta_i) \Big\}.$$

Matching coefficients on Δ_i: $r\mu_i = B_i - \mu_i \lambda_i$, i.e. $\mu_i = B_i/(r + \lambda_i)$; the constant v_0 collects the maximized $\sum_i \{\mu_i x_i^* - c(x_i^*)\}$ term. The maximand is concave, so $x_i^* = c'^{-1}(\mu_i)$ (interior) or, with linear cost $c_i x_i$ and a budget, the bang-bang rule in μ_i/c_i. Since flow and dynamics are linear and the gap is bounded under $\lambda_i > 0$, the linear V is the value function by a standard verification argument.

(ii) Investment is horizon-directed iff $\mu_H/c_H > \mu_F/c_F$, i.e. iff $B_F/B_H < \rho^* = (r +$

$\lambda_F)c_F/\{(r+\lambda_H)c_H\}$. Along a horizon push q_H increases while q_F is constant, and

$$\frac{d}{dq_H}\ln\frac{B_F}{B_H} = \frac{C(q)}{B_F} - \frac{\partial B_H/\partial q_H}{B_H} \;>\; 0$$

under the stated conditions $C > 0$ and $\partial B_H/\partial q_H \le 0$, using (4). If the ratio eventually exceeds ρ^*, it crosses it exactly once (strict monotonicity), at a finite frontier position and hence—investment rates being bounded below on the horizon phase—at a finite time T^*; before the crossing $\mu_H/c_H > \mu_F/c_F$ and after it the ranking flips.

(iii) $\partial\rho^*/\partial\lambda_H = -(r+\lambda_F)c_F/\{(r+\lambda_H)^2 c_H\} < 0$ and $\partial\rho^*/\partial\lambda_F = c_F/\{(r+\lambda_H)c_H\} > 0$. With B_F/B_H strictly increasing along the push, a lower threshold is crossed earlier, so T^* falls with λ_H and rises with λ_F. □

A.6 Proof of Theorem 4

Payoffs: same direction k gives σV_k each; split gives each its own V_{d_i}. Against an opponent on F, playing F yields σV_F and playing H yields V_H; against an opponent on H, playing F yields $V_F \ge \sigma V_H$ always (as $V_F \ge V_H > \sigma V_H$). *(i)* If $\sigma V_F \ge V_H$, F is a best reply to both opponent actions, strictly dominant when strict, and (F,F) is the unique equilibrium. *(ii)* If $\sigma V_F < V_H$, then H is the unique best reply to F and F to H: the pure equilibria are (F,H) and (H,F). Indifference for the mixed equilibrium, with $p = \Pr(\text{opponent plays } F)$: $p\,\sigma V_F + (1-p)V_F = p\,V_H + (1-p)\sigma V_H$, giving

$$p^* = \frac{V_F - \sigma V_H}{(V_F - \sigma V_H) + (V_H - \sigma V_F)} = \frac{\rho - \sigma}{(1-\sigma)(\rho+1)}, \qquad \rho = \frac{V_F}{V_H},$$

which lies in $(0,1)$ exactly on the stated region, and $dp^*/d\rho = (1+\sigma)/\{(1-\sigma)(\rho+1)^2\} > 0$. *(iii)* The herding condition $\sigma \ge V_H/V_F = 1/\rho$ relaxes as ρ rises; ρ is increasing in $\mu_F/\mu_H = \{B_F/B_H\}\{(r+\lambda_H)/(r+\lambda_F)\}$, which is increasing in λ_H and decreasing in λ_F. □

A.7 Proof of Theorem 5 and Corollary 1

The planner's flow surplus is $W(q^L(t))$: served value depends on the best available capability, and reallocation of margins between frontier and fringe is a transfer. A unit of direction-i capability raises the flow by B_i permanently, so its present value is B_i/r. The market values it at $\mu_i = B_i/(r+\lambda_i)$ by Theorem 3(i). The ratio of relative valuations is

$$\frac{\mu_H/\mu_F}{(B_H/r)/(B_F/r)} = \frac{r+\lambda_F}{r+\lambda_H},$$

which is below one iff $\lambda_H > \lambda_F$, proving (i); (ii) follows by replacing the planner's B_F with $\theta_s B_F$ and comparing θ_s with $(r+\lambda_H)/(r+\lambda_F)$. For the corollary, quadratic costs give $x_F^* = \mu_F$;

under liability the private reliability boundary value is θB_F, so

$$x_F^* = \frac{\theta B_F}{r + \lambda_F}, \qquad \Delta_F^* = \frac{x_F^*}{\lambda_F} = \frac{\theta B_F}{(r + \lambda_F)\lambda_F},$$

whence $\partial x_F^*/\partial\theta > 0$ and $\partial\Delta_F^*/\partial\theta > 0$. Differentiating in λ_F: $\partial\Delta_F^*/\partial\lambda_F = -\theta B_F(r + 2\lambda_F)/\{(r + \lambda_F)^2\lambda_F^2\} < 0$ and $\partial x_F^*/\partial\lambda_F < 0$; the fringe's steady-state reliability is $q_F^L - \Delta_F^*$, which rises toward the frontier as λ_F rises. □

B Machine Verification

Every formal claim in the paper is checked in a companion script, `math_check_v3.py`, which verifies symbolically (SymPy) and numerically: the threshold identity and sandwich (7); the two-term error bound and the $1/(1 + \ell)$ slope of Lemma 1(i); the Gumbel limit and its $O(n^{-1/(1+\ell)})$ rate; the parameter-free $\ln 0.5/\ln 0.8 \approx 3.106$ ratio and its invariance to per-stage success and retries; the representation logic of Theorem 1 on explicit task sets; the derivative, cross-derivative, and path identities (3)–(5); the Bertrand margin and profit–gap identity and saturation limit of Lemma 2; the χ^2/KL bounds and sample-floor constants of Theorem 2; the HJB coefficients, rotation single-crossing, and threshold comparative statics of Theorem 3 on a fully specified two-family task economy; the equilibrium regions, mixed-strategy formula, and monotonicity of Theorem 4; the wedge algebra of Theorem 5; the steady-state policy derivatives of Corollary 1; and the calibration arithmetic of Section 9. The script reports 0 failures.

References

Acemoglu, Daron. 2002. "Directed Technical Change." *Review of Economic Studies* 69(4): 781–809.

Acemoglu, Daron. 2007. "Equilibrium Bias of Technology." *Econometrica* 75(5): 1371–1409.

Acemoglu, Daron. 2025. "The Simple Macroeconomics of AI." *Economic Policy* 40(121): 13–58.

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics*, Vol. 4B, 1043–1171. Elsevier.

Acemoglu, Daron, and Pascual Restrepo. 2018. "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment." *American Economic Review* 108(6): 1488–1542.

Aghion, Philippe, Christopher Harris, Peter Howitt, and John Vickers. 2001. "Competition, Imitation and Growth with Step-by-Step Innovation." *Review of Economic Studies* 68(3): 467–492.

Anton, James J., and Dennis A. Yao. 1994. "Expropriation and Inventions: Appropriable Rents in the Absence of Property Rights." *American Economic Review* 84(1): 190–209.

Anton, James J., and Dennis A. Yao. 2002. "The Sale of Ideas: Strategic Disclosure, Property Rights, and Contracting." *Review of Economic Studies* 69(3): 513–531.

Arrow, Kenneth J. 1962. "Economic Welfare and the Allocation of Resources for Invention." In *The Rate and Direction of Inventive Activity*, 609–626. Princeton University Press.

Bloom, Nicholas, Charles I. Jones, John Van Reenen, and Michael Webb. 2020. "Are Ideas Getting Harder to Find?" *American Economic Review* 110(4): 1104–1144.

Bresnahan, Timothy F., and Manuel Trajtenberg. 1995. "General Purpose Technologies: 'Engines of Growth'?" *Journal of Econometrics* 65(1): 83–108.

Bryan, Kevin A., and Jorge Lemus. 2017. "The Direction of Innovation." *Journal of Economic Theory* 172: 247–272.

Budish, Eric, Benjamin N. Roin, and Heidi Williams. 2015. "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials." *American Economic Review* 105(7): 2044–2085.

Callander, Steven. 2011. "Searching and Learning by Trial and Error." *American Economic Review* 101(6): 2277–2308.

Dasgupta, Partha, and Eric Maskin. 1987. "The Simple Economics of Research Portfolios." *Economic Journal* 97(387): 581–595.

Gilbert, Richard J., and David M. G. Newbery. 1982. "Preemptive Patenting and the Persistence of Monopoly." *American Economic Review* 72(3): 514–526.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. "Distilling the Knowledge in a Neural Network." arXiv:1503.02531.

Hsieh, Cheng-Yu, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. "Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes." arXiv:2305.02301.

Jones, Charles I. 2011. "Intermediate Goods and Weak Links in the Theory of Economic Development." *American Economic Journal: Macroeconomics* 3(2): 1–28.

Korinek, Anton, and Jai Vipra. 2025. "Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence." *Economic Policy* 40(121): 225–256.

Kremer, Michael. 1993. "The O-Ring Theory of Economic Development." *Quarterly Journal of Economics* 108(3): 551–575.

Kwa, Thomas, Ben West, Joel Becker, et al. 2025. "Measuring AI Ability to Complete Long Tasks." arXiv:2503.14499. METR.

Mussa, Michael, and Sherwin Rosen. 1978. "Monopoly and Product Quality." *Journal of Economic Theory* 18(2): 301–317.

Ord, Toby. 2025. "Is There a Half-Life for the Success Rates of AI Agents?" arXiv:2505.05115.

Reinganum, Jennifer F. 1983. "Uncertain Innovation and the Persistence of Monopoly." *American Economic Review* 73(4): 741–748.

Shaked, Avner, and John Sutton. 1982. "Relaxing Price Competition through Product Differentiation." *Review of Economic Studies* 49(1): 3–13.

Teece, David J. 1986. "Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing and Public Policy." *Research Policy* 15(6): 285–305.