

The Asymmetry Engine, Chapter 11: The loop that closes on itself

Martin Roark

Chapter 11 of *The Asymmetry Engine: How AI Closes the Gaps, and What Stays Scarce*, October 2026. The online edition, with the interactive figures and the links between chapters, is at <https://future-seems-so-good.com/blog/the-asymmetry-engine#recursion>. License: [CC BY 4.0](#).

11:1 Automated AI research is the one closure that raises intelligence supply itself. Every other gap closes at a rate proportional to the intelligence aimed at it; the research gap is the only one whose closing adds to that intelligence, so it feeds back on every closing rate at once. Whether the loop runs away is set by three numbers: the returns to research, how well automated thinking substitutes for experiment compute, and how independent the automated researchers are of one another.

11:2 The best estimates sit just above the knife-edge between a loop that decelerates and one that explodes. Compute is the homeostat: on fixed hardware the runaway is a transient, and on growing hardware the speed of software is pinned to the build-out (Proposition 11.2). Runaway means crossing an ignition threshold nobody outside a lab can locate (Proposition 11.3). A delayed brake holds only a loop slower than its lag, and the first year of full automation depends more on the launch speed than on the exponent (Experiment 11.1).

11.1 The loop, stated

11:3 The last line of the master equation, $\dot{I} = \eta \lambda_{\text{res}} G_{\text{res}}$, is the loop. Raising research capability raises I and with it every λ_k ; raising any other capability moves one closing rate. Four terms of (0.2) are at stake: η , how research output becomes intelligence supply; v_{res} , how cheaply a research result can be checked; a_{res} , how much autonomy alignment licenses; and φ_{res} , the inertia of experiments.

11:4 I. J. Good stated the loop in 1965: since "the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines," the first one would be "the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control" ([Good 1965](#)). Bostrom wrote its kinetics as optimization power over recalcitrance; Yudkowsky expected such a loop to "either peter out rapidly ... or else go FOOM" ([2008](#)); Christiano defined the slow alternative, world output doubling over four years before it ever doubles in one ([2018](#)).

11:5

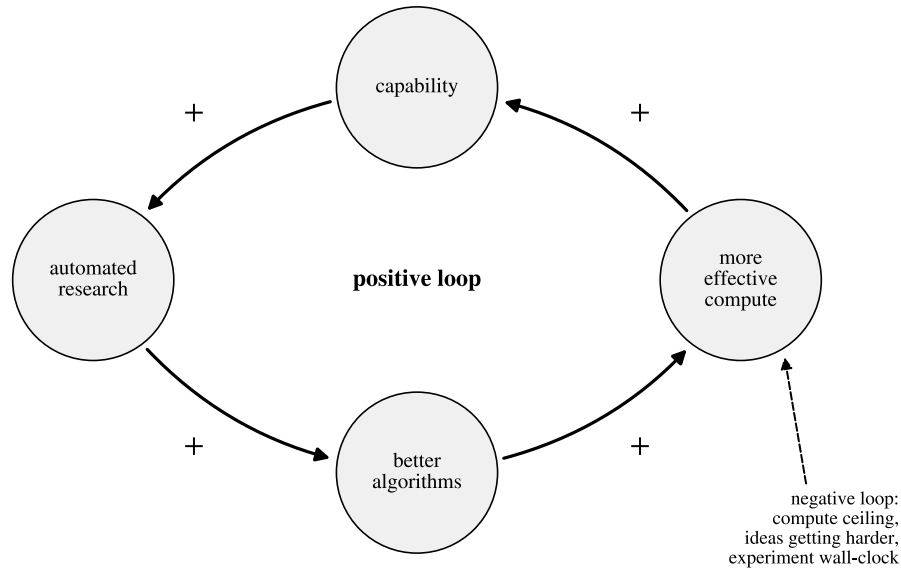


Figure 11.1. The research loop: capability, automated research, better algorithms and more effective compute joined in a circle by four positive links, with a negative loop (the compute ceiling, ideas getting harder, the wall-clock time of experiments) acting on effective compute. The diagram gives the loop's signs, not its strength: with research input Cobb–Douglas in labor and compute, the returns to research decide between polynomial, exponential and finite-time growth (Proposition 11.1), and the compute ceiling is the homeostat that bounds the runaway even when those returns exceed one (Proposition 11.2).

11:6 Two forces meet in one number: ideas get harder to find, each unit of progress raising the cost of the next even while measured progress speeds up (§2.7), and automated researchers multiply research input.

11:7

Proposition 11.1 (The singularity condition). Let S be software efficiency and E research input, with Jones's idea production $\dot{S}/S = g_J S^{-\beta} E^\theta$, where β measures how much harder ideas get as software improves and θ how much parallel researchers step on one another's toes. Research input is Cobb–Douglas in automated labor L and experiment compute C , with labor's share s_L . Under full automation on fixed compute, labor is proportional to software:

$$E \propto L^{s_L} C^{1-s_L}, \quad L \propto S \implies \dot{S} \propto S^{1+\theta s_L - \beta}.$$

With the **returns to research** $r = \theta s_L / \beta$, growth is polynomial, exponential or finite-time as r is below, at or above 1, and each doubling of software takes $2^{\theta s_L (1/r - 1)}$ times as long as the last. The exponent on S is $1 + \theta s_L (1 - 1/r)$, never r itself. The blow-up for $r > 1$ belongs to this Cobb–Douglas law: with labor and compute complements (Proposition 11.2), or with capability that decays (Proposition 11.3), $r > 1$ alone does not imply it.

11:8 With constants absorbed into g_J and $q = 1 + \theta s_L - \beta$ (local), the law $\dot{S} = g_J S^q$ reaches infinity at a finite time when $q > 1$:

11:9

$$S(t) = S_0 \left(1 - \frac{t}{t^*}\right)^{-1/(q-1)}, \quad t^* = \frac{S_0^{1-q}}{g_J (q-1)} \quad (11.1)$$

11:10 This is the finite-time blow-up von Foerster fitted to world population in 1960 (§0.1), and the exponent sets its horizon. With Forethought's own penalty for parallel research, $\theta s_L \approx 0.3$, and a first doubling of one month, returns of 1.2, 1.4 and 3 sum all the doublings to 29.4, 17.3 and 7.7 months, and at 0.7 each doubling takes 9% longer than the last. Assuming that every doubling of the workforce doubles research ($\theta s_L = 1$) makes each horizon about three times shorter.

▼ *Derivation: The cumulative-effort form*

Eth and Davidson define the returns through cumulative effort, $S \propto R_{\text{cum}}^r$, with R_{cum} the cumulative research input (local): r "gives the number of times software doubles for each time the cumulative work on software R&D doubles" (Forethought 2025). If automated researchers add effort one-for-one with software, $\dot{R}_{\text{cum}} \propto R_{\text{cum}}^r$, and for $r > 1$ software follows $S \propto (t^* - t)^{-r/(r-1)}$, the profile of $\dot{S} \propto S^{2-1/r}$: the proposition's law with $\theta s_L = 1$. If the first doubling takes T_{first} (local), each later one takes $2^{\theta s_L(1/r-1)}$ times the one before, and for $r > 1$ all of them sum to

$$T_{\text{tot}} = \frac{T_{\text{first}}}{1 - 2^{\theta s_L(1/r-1)}}$$

(T_{tot} local): the horizons above at $\theta s_L = 0.3$, and 9.2, 5.6 and 2.7 months at $\theta s_L = 1$.

11.2 The contested number

11:11 Estimates of the returns to research straddle 1 once they are deflated for compute's share of research. Deflation leaves the knife-edge at $r = 1$ and moves each estimate's distance from it, which sets the speed.

11:12 Measured against research effort alone, the returns sit above 1: Ho and Whitfill find 1.26 for computer vision, 1.20 for reinforcement learning and 1.89 for language models (Epoch 2025). Other fields span the threshold, from 0.25–0.32 for the US economy and about 0.83 for the chess engine Stockfish (standard error 0.15) (Epoch 2024) to 1.1 for linear programming, 1.6 for sample efficiency in reinforcement learning and 3.5 for SAT solvers (Eth & Davidson).

11:13 Research runs on compute as well as ideas. If compute is about two-thirds of research inputs and automation multiplies only labor, Ho and Whitfill's estimates "should be cut by a factor of three, which puts them all below 1": about 0.40–0.63. Davidson and Houlden's median of 1.2, log-uniform over 0.4–3.6, is already deflated (Forethought 2025). The two camps disagree on the sign of $r - 1$ after deflation and agree before it.

11:14 At Davidson and Houlden's $\theta = 0.6$ and $s_L = 0.5$, the median gives an exponent of 1.05, and the prior spans 0.55 to 1.22. Each doubling is then 3% shorter than the last at the median, 14% shorter at 3.6 and 37% longer at 0.4: gentle acceleration in the middle, an explosion only in the upper tail. Their bottom line is about 60% that an explosion compresses more than three years of progress into one, about 20% for more than ten, and no thirtyfold speed-up without a new paradigm (Forethought 2025).

11:15 Substitutability is the hinge. On a panel of four labs over 2014–2024, Whitfill and Wu estimate the elasticity of substitution between research compute and labor at $\varepsilon_s = 2.58$ (standard error 0.34), substitutes, and at

—0.10 (0.18), "statistically indistinguishable from zero", once frontier-scale experiments are modeled; they prefer the second ([Whitfill & Wu 2025](#)). Erdil and Barnett take manufacturing's 0.7 and find that a software-only singularity would "fizzle out after less than an order of magnitude of improvement in efficiency" ([Epoch 2025](#)); Davidson puts the elasticity most likely between 0.83 and 1, because fast-thinking AIs could reorganize research around compute ([Forethought 2025](#)). A singularity and a very productive decade differ in whether a cleverer researcher can stand in for a bigger experiment, and as of October 2026 the two best estimates, from the same authors on the same data, fall on opposite sides of 1.

11:16 Forces push both ways. Toward substitutes: efficiency gains cheapen each experiment, small runs extrapolate, as GPT-4 was predicted from runs with "<1/1,000th as much computing power", and automated researchers can design fewer, more informative experiments ([Eth & Davidson](#)). Toward complements: much measured progress comes from a few scale-dependent changes ([Ho 2026](#)); perhaps 99% of historical compute-equivalent gains came from compute-dependent innovations ([Josephson 2025](#)); and Trammell adds the "parallelization technology" needed to "divide, execute, coordinate, and integrate" many researchers' work ([Trammell 2026](#)), which sequential research lacks (§9.7).

11.3 Compute is the homeostat

11:17 Compute is the homeostat of the research loop. A million automated researchers with no new chips queue for the same experiments as a few thousand, and the steadiest regularity in the evidence is the signature: progress multipliers fall far below labor multipliers.

11:18

Source	Research labor added	Progress gained
Aschenbrenner, 2024	a millionfold research effort	"a 10x acceleration"
AI 2027, 2025	200,000 automated coders	"only" 4×
Anthropic, 2026	about 4× staff output	"below 2×"; 2× would take uplift "roughly an order of magnitude larger"
METR, 2026	8× code, about 2.8× per researcher	2× overall needs about 3.5× per researcher
OpenAI, 2026	3.1 agent-workdays per human workday	"likely won't keep pace with these specific metrics"

11:19 The law of motion, the research line of (0.2) made dynamical, builds that complementarity into the loop. Automated labor equals software, $L = S$, because better software runs more and faster agents on the same fleet ((10.1) converts gigawatts into agents); experiment compute grows as $C = C_0 e^{g_C t}$; the **launch speed** ξ is the research speed-up at the moment of full automation, in multiples of $g_0 = \ln 3$ a year, the 2020–24 pace of software progress; and $S_{\max}$ stands for physical limits. As $\varepsilon_s \rightarrow 1$ the bracket becomes Cobb–Douglas and the law becomes Davidson and Houlden's model; below 1, labor and compute are complements:

11:20

$$\frac{d \ln S}{dt} = \xi g_0 \left[s_L L^{\frac{\varepsilon_s - 1}{\varepsilon_s}} + (1 - s_L) C^{\frac{\varepsilon_s - 1}{\varepsilon_s}} \right]^{\frac{\theta \varepsilon_s}{\varepsilon_s - 1}} S^{-\beta} \left(1 - \frac{\ln S}{\ln S_{\max}} \right), \quad r = \frac{\theta s_L}{\beta} \quad (11.2)$$

Proposition 11.2 (Compute is the homeostat). In the law of motion (11.2), away from the physical limit: (i) With fixed compute, the speed of progress $d \ln S / dt$ first rises if and only if $r > 1$. (ii) If labor and compute are complements ($\varepsilon_s < 1$), the rise ends when

$$\ln S = \frac{\varepsilon_s}{1 - \varepsilon_s} \ln \frac{r - s_L}{1 - s_L},$$

and the speed falls afterward: the runaway is a transient. At $s_L = \frac{1}{2}$, infinite labor raises research input at most $2^{\varepsilon_s/(1-\varepsilon_s)}$ -fold: 2× at $\varepsilon_s = 0.5$, 5.5× at 0.71, 30× at 0.83. (iii) With compute growing at g_C and labor abundant, compute limits research input but not software itself, and the speed converges to $(\theta/\beta) g_C = (r/s_L) g_C$, about 2.4 times compute's growth at the median $r = 1.2$. (iv) If compute also caps software itself, through a ceiling on S that grows at g_C , software grows no faster than g_C in the long run, and with complements a loop with $r > s_L$ rides the ceiling at the build-out's rate while one with $r < s_L$ falls behind it. Compute is therefore the **homeostat**, the brake that bounds the runaway: while it binds, hardware sets the takeoff rate, and only software that cheapens experiments lets the loop outrun it.

▼ *Proof*

With $C = 1$, $L = S$ and no physical limit,

$$\frac{d}{d \ln S} \ln \frac{d \ln S}{dt} = \theta \hat{s}_L(S) - \beta, \quad \hat{s}_L(S) = \frac{s_L S^{(\varepsilon_s-1)/\varepsilon_s}}{s_L S^{(\varepsilon_s-1)/\varepsilon_s} + 1 - s_L},$$

where $\hat{s}_L(S)$ (local) is labor's share of research input at progress S , equal to s_L at the start. At $S = 1$ the derivative is $\theta s_L - \beta = \beta(r - 1)$, which proves (i). For $\varepsilon_s < 1$, $\hat{s}_L$ falls monotonically toward zero, so the derivative changes sign once, where $\hat{s}_L = \beta/\theta = s_L/r$; solving for S gives (ii). As $L \rightarrow \infty$ research input tends to $(1 - s_L)^{\varepsilon_s/(\varepsilon_s-1)} C$, the stated factor at $s_L = \frac{1}{2}$. For (iii), when $L \gg C$ the speed is proportional to $C^\theta S^{-\beta}$, and a constant speed needs $\theta g_C = \beta d \ln S / dt$. For (iv), S cannot pass the ceiling, so its growth cannot beat g_C . On the ceiling S grows at g_C and labor at least as fast, so with complements research input grows at g_C and the factor $E^\theta S^{-\beta}$ of the speed at $(\theta - \beta) g_C$: positive when $r > s_L$, so the loop presses on the ceiling and rides it, and negative when $r < s_L$, so it cannot stay there. Figure 11.2 draws the reduced law $\dot{S} = g_J S^q$ under the same ceiling, with compute entering only there; in it the ratio to the ceiling tends to 1 when $r > 1$, to $1 - g_C/g_J$ at $r = 1$ (§C.4) and to zero when $r < 1$. ■

11:22 At the median the loop barely accelerates before compute binds: at $r = 1.2$, $\theta = 0.6$ and $s_L = 0.5$, the speed rises 2–9% over less than one and a half orders of magnitude, fast exponential progress. The runaway's length and the logarithm of its height scale as $\varepsilon_s/(1 - \varepsilon_s)$, the homeostat's gain, so even $r = 3$ against strong complements ($\varepsilon_s = 0.71$) gains 43% over 1.7 orders of magnitude and then decelerates. The homeostat follows from complementarity alone, and Experiment 11.1 measures it.

11:23 Clause (iii) ties the loop to the gigawatts: once automated labor is abundant, the long-run speed of software is compute's growth times r/s_L , Jones's semi-endogenous growth with compute in the place of population (Jones 1995). If experiment compute grows 2.5–4× a year, software grows about 9–28× a year until physical limits bind; under the harder ceiling of clause (iv), which Figure 11.2 draws, it grows only as fast as compute. The gigawatts

set the homeostat's setpoint, and capital spending turns the dial (§16.1). Epoch's best guess at today's pace, about 10× a year (80% interval 2–50×), sits inside that band (Ho 2026), so the level cannot tell the regimes apart.

11:24

Forecast 11.1 (The setpoint test). Software progress per unit of compute growth in 2027 is less than twice its 2026 value, as a compute-bound loop predicts.

Horizon 2028-06-30

Probability 60%

Check Read Epoch AI's estimates, or lab disclosures, of yearly software-efficiency growth and of experiment or training compute growth for 2026 and 2027, the latest published by the horizon, and take each year's ratio of the logarithms of the two growth factors. The forecast holds if the 2027 ratio is less than twice the 2026 ratio; it fails if it is twice or more, or if no estimate for 2027 is published.

11:25

Interactive in the online edition: <https://future-seems-so-good.com/blog/the-asymmetry-engine#fig:rsi>

Figure 11.2 (interactive). The research loop, as speed against progress on the left and progress against time on the right. Slide r through 1 and watch polynomial growth turn exponential and then blow up at a finite time. Then switch on the cap, let compute grow at g_C , and watch the blow-up bend into steady growth at the build-out's rate. Compute enters only through the cap here, so loops with $r < 1$ fall behind it; when compute also feeds research, every loop with $r > s_L$ keeps up (Proposition 11.2).

11:26

Proposition 11.3 (The ignition threshold). With a decay rate d_S of capability that is not reinvested (models are superseded, environments saturate, verifiers are gamed), the law $\dot{S} = g_J S^{1+\theta s_L - \beta} - d_S S$ with $\theta s_L > \beta$ has an unstable equilibrium, the **ignition threshold**

$$S_{\text{ign}} = \left(\frac{d_S}{g_J} \right)^{1/(\theta s_L - \beta)}.$$

Below it, capability decays back toward homeostasis. Above it, capability blows up at

$$t^* = - \frac{\ln(1 - d_S S_0^{-(\theta s_L - \beta)} / g_J)}{(\theta s_L - \beta) d_S},$$

which tends to the blow-up time of (11.1) as $d_S \rightarrow 0$. Runaway means crossing the ignition threshold; $r > 1$ is necessary and not sufficient.

▼ *Proof*

The quantity $S^{-(\theta s_L - \beta)}$ obeys the linear equation $\frac{d}{dt} S^{-(\theta s_L - \beta)} = (\theta s_L - \beta) (d_S S^{-(\theta s_L - \beta)} - g_J)$, whose equilibrium g_J/d_S repels. Capability blows up when the quantity reaches zero, which happens, at t^* , if and only if it starts below g_J/d_S , that is, if and only if $S_0 > S_{\text{ign}}$. A fourth-order Runge–Kutta integration confirms the blow-up time. ■

11:27 Positive feedback runs away only past this threshold: escape from homeostasis is the crossing of an unstable equilibrium, like a reactor going critical, and near the knife-edge the threshold is hypersensitive. At the median exponent of 1.05, $\ln S_{\text{ign}} = 20 \ln(d_S/g_J)$, so a 5% change in the ratio of decay to research productivity moves the threshold by a factor of 2.7. Nobody outside a lab can know which side the loop is on, so the observable that matters is the sign of the change in doubling times (Forecast 11.5).

11:28 The same clauses sort the geopolitics. A lab can import software through distillation; it cannot import the homeostat's setpoint, which is counted in gigawatts (§12.7).

11.4 What the Monte Carlo says

11:29 Sampling the law of motion over the published priors reproduces the literature's tail and shows that the first year turns on the launch speed.

11:30

Experiment 11.1 (The research loop, sampled).

Setup: the law of motion (11.2), integrated over four years for 12,000 draws per scenario.

Parameters: r log-uniform on 0.4–3.6 (median 1.2); θ uniform on 0.4–0.8 with $s_L = 0.5$; launch speed ξ log-uniform on 2–32 $\times$ (median 8 $\times$); ε_s mixed (0.67–1), strong complements (0.67–0.83) or Cobb–Douglas (1); physical limits 6–16 orders of magnitude away; compute fixed, or growing 2.5–4 $\times$ a year.

Result: with fixed compute and mixed ε_s , a 20% chance that the first year of full automation compresses more than ten years of 2020–24-pace progress (79% for more than three; median 5.3 years, 2.5 orders of magnitude); 35% under Cobb–Douglas and 15% with strong complements; with compute growing, six orders of magnitude within two years become 37% likely instead of 23% (§B.7); year-one progress tracks the launch speed (Spearman 0.80) more than r (0.52).

Shows: the tail matches Davidson and Houlden; the bulk is an upper envelope, because ramp-up and retraining delays are omitted.

11:31

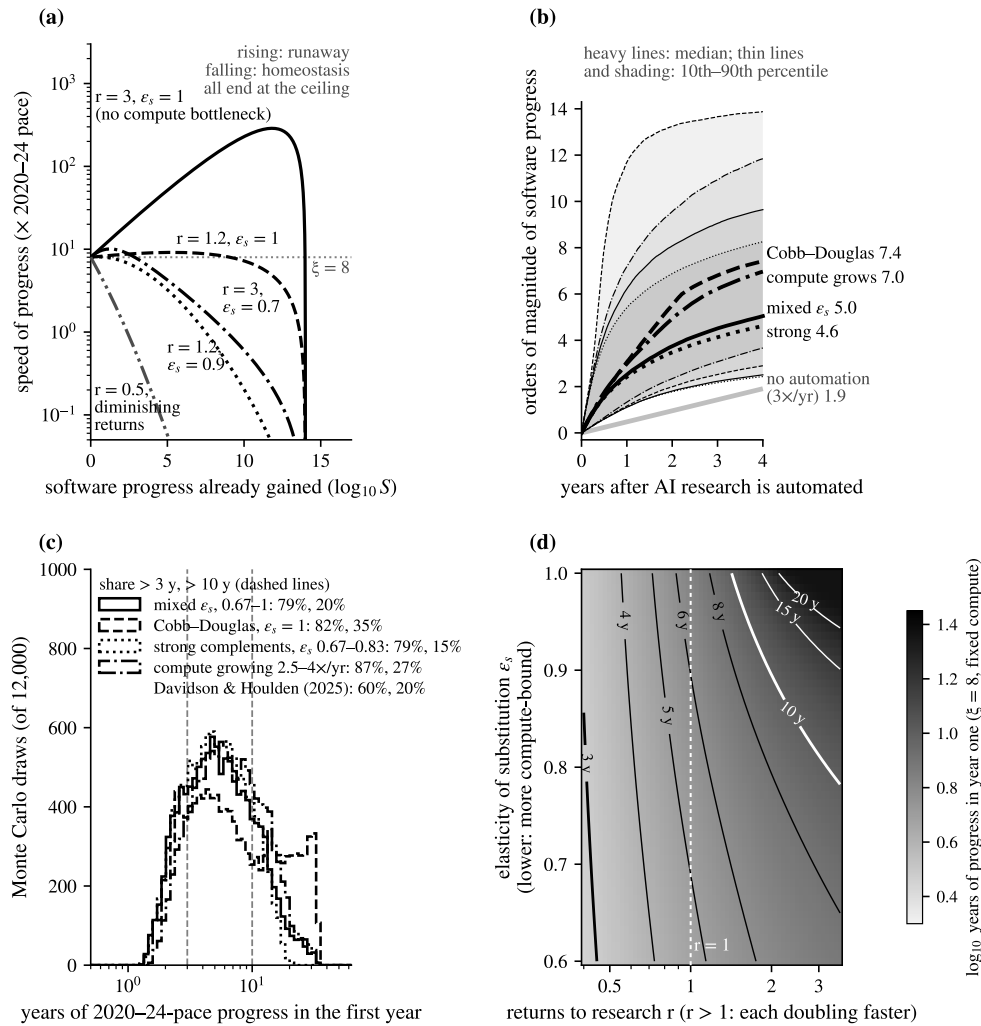


Figure 11.3. The sampled loop: (a) speed of progress, in multiples of the 2020–24 pace, against progress made at $\xi = 8$, for $r = 3$ at ε_s of 1 and 0.7, $r = 1.2$ at 1 and 0.9, and $r = 0.5$ at 0.9; (b) median trajectories with 10–90% bands; (c) year-one progress by scenario, with the shares above three and ten years against Davidson and Houlden's 60% and 20%; (d) year-one progress over r and ε_s at $\xi = 8$. Rising curves in (a) are positive feedback winning, up to about 290 times the 2020–24 pace at $r = 3$ and $\varepsilon_s = 1$, and falling ones homeostasis; the ten-year contour in (d) fills only the corner of high r and high ε_s .

11:32 Forethought and Epoch disagree about where we sit in panel (d): at an elasticity between 0.83 and 1 with returns of 1.2, fast exponential progress with an explosive tail, or near 0.7 with deflated returns below 1, where the first year still brings about 4.5 years of progress and a software-only explosion fizzles afterward, each doubling slower than the last.

11:33 Year-one progress has a rank correlation of 0.80 with the launch speed and 0.52 with the returns to research, against 0.14 with the elasticity and none with θ . For the first year, how good and how numerous the automated researchers are at launch matters more than the exponent that decides an explosion. The launch speed is a product of four factors: the agent count ((10.1)), the quality per agent, a discount for correlated clones (Proposition 11.4) and a cap from verification throughput ((9.2)); the experiment folds the last two into ξ and θ .

11:34 Compute growing 2.5–4× a year, of which gigawatts supply about 2–2.4× (Experiment 10.1) and gains per watt the rest, lifts the chance of more than ten years in the first year from 20% to 27% (§B.7): compute growth

moves the setpoint and turns a transient into a sustained climb. Capital spending is part of the loop (§16.1): the labs already send their marginal megawatt increasingly to research, and the two largest may hold most of the world's compute by 2028 ([Dylan Patel, August 2026](#)). A full-stack explosion closes three loops, software, chip technology and chip production, with about 13, 6 and 5 orders of magnitude of headroom ([Forethought 2025](#)).

11:35 The model has one sector, no retraining lag, no hardware feedback and a stylized limit, and it counts years of progress against a 3×-a-year baseline. Requiring training time before software improvements apply "reduces the chance of very fast takeoffs" ([AI Futures, August 2026](#)), so the bulk is an upper envelope; §B.7 has the method and full tables.

11.5 The fuel in 2026

11:36 In 2026 the loop is closing with people still inside it: execution uplift is large, task horizons double every three to four months, field uplift lies between about 1.2× and 4×, and judgment uplift is thin.

11:37 On a fixed training-code optimization task, Claude's speed-up over the baseline went from about 3× (Opus 4, May 2025) to about 52× (Mythos Preview, April 2026), where a skilled human reaches 4× in four to eight hours; Anthropic warns that the figure "should not be read as a real-world training speedup." More than 80% of the code Anthropic merges is now written by Claude ([Anthropic, June 2026](#)). Nine parallel Claude agents recovered 97% of a weak-to-strong supervision gap in 800 agent-hours for about \$18,000, against 23% for two researchers in a week, a result that "didn't transfer cleanly to production-scale models" ([Anthropic, April 2026](#)).

11:38 Swarms find methods where a verifier exists. AlphaEvolve found a kernel 23% faster, cutting Gemini's training time by 1%, and a scheduling heuristic that recovers 0.7% of Google's compute, and it improved "the large language models underlying AlphaEvolve itself" ([Google DeepMind 2025](#)). A Cursor and NVIDIA swarm sped up 235 CUDA problems by 38% (geometric mean) in three weeks ([Cursor 2026](#)), and the Darwin Gödel Machine went from 20% to 50% on SWE-bench by rewriting its scaffold with frozen weights ([Zhang et al. 2025](#)). No public case shows a swarm discovering an architecture. Every success has a cheap verifier, and the verifier must be nearly perfect, as Anthropic's compiler-writing agents showed (9:14).

11:39 METR's 50%-success task horizon has doubled every 130.8 days for models since 2023 and every 88.6 days since 2024 ([METR, January 2026](#)); Mythos Preview measured "at least 16hrs (95% CI 8.5hrs to 55hrs)", beyond which METR calls its suite unreliable ([METR](#)).

11:40 Field uplift is bracketed rather than measured. METR's 2025 trial found experienced open-source developers 19% slower with early-2025 tools while they believed AI had sped them up by 20% ([METR 2025](#)); its follow-up finds tasks taking about 18% less time for returning developers (CI −38% to +9%), "likely... a lower-bound", because 30–50% of developers withheld tasks they would not do without AI ([METR, February 2026](#)). Anthropic's staff survey puts the geometric mean "on the order of 4×" ([system card](#)). The true uplift lies between about 1.2×, biased down by selection, and 4×, biased up by self-report. It sets the labor term of the law of motion, nobody outside the labs can measure it, and a regulator that cannot measure a loop cannot model it (§4.6).

11:41 Judgment uplift is thin. Anthropic reports that "large performance gaps persist when it comes to Claude exercising judgement in choosing goals in both engineering and research" ([Anthropic, June 2026](#)), and of ideas that "vault the field forward" Jack Clark says "we don't see that yet" ([Import AI 460](#)). Execution has a large

verification asymmetry and falls first. Research taste, choosing which experiment is worth running, has $\alpha \approx 1$, because verifying a research direction takes about as long as pursuing it (Definition 2.1).

11:42 OpenAI reached its self-set "automated research intern" goal on 6 September 2026: "a system that can carry out well-defined research tasks under human direction, including tasks that would take a skilled researcher a few days." By mid-August its research organization used 3.1 agent-workdays per human workday, its median researcher more than \$600 of inference a day at API prices. The figures are self-reported, most successful tasks of four to eight hours still needed a human intervention (8:11), "people still set our research priorities," and the target for "an automated AI researcher" is March 2028 ([OpenAI](#)). OpenAI's own summary is that "fully autonomous RSI is not happening today" ([via The Next Web](#)), and the one external audit agrees: METR judged on 22 September that Claude Opus 5.5 is "unlikely to be able to fully automate AI R&D", with an acceleration of about 1.5× ([METR](#)). Uptime is not autonomy (§8.3).

11:43 A researcher's output is throughput times taste. Copying a model multiplies throughput at a cost the gigawatt arithmetic fixes ((10.1)); it multiplies taste only if taste is in the weights. Ten million clones of a model with a top researcher's throughput and an intern's taste are ten million interns: valuable, and short of a singularity.

11.6 Ten million researchers

11:44 Ten million clones of one checkpoint are worth far fewer independent researchers, and ten million running at ten thousand tokens a second each is a picture a hardware generation early. The count is likely, the speed is not, and the quality is the open question.

11:45 The unit is Alec Radford, lead author of GPT and GPT-2 and first author of CLIP and Whisper. Leopold Aschenbrenner made him the unit of automated research in 2024: "Imagine an automated Alec Radford—imagine 100 million automated Alec Radfords." He saw the constraint, "limited compute for experiments will be the bottleneck," and found it "hard to believe that the 100 million Alec Radfords couldn't increase the marginal product of experiment compute by at least 10x" ([Situational Awareness](#)).

11:46 Such a swarm would work on the training stack, where every task has a verifier (measured kernel speed, loss curves, held-out evaluations) and the loop moves fastest. The code is undisclosed, but a full-stack pipeline can be small: Karpathy's nanochat is about 8,000 lines ([Willison](#)). Code that size admits few simultaneous editors (Proposition 9.1), so ten million researchers would parallelize over experiments, which cost compute.

11:47 Clones without ego fuse two properties, aligned goals and correlated errors (7:43). Trammell puts it exactly: "even if every instance of Claude Fable 5 were in every sense as smart as Einstein, their biases are highly correlated across instances, and a hundred Einsteins might have taken longer to advance quantum mechanics than an Einstein and a Bohr." One mind copied many times loses its ego and its variety together, and only variety absorbs variety ((4.1)).

11:48

Proposition 11.4 (Clones as a jury and as a search). Let the quality of agent j 's best idea be an increasing function of $\sqrt{\rho} Z_0 + \sqrt{1 - \rho} Z_j$, with a shared standard-normal factor Z_0 and independent factors Z_j . As a jury on a binary question, N clones tend to the Condorcet ceiling of Proposition 7.3 however large N grows. As a search behind a sound verifier that keeps the best idea, they have no ceiling but reach only as deep into the tail as about $N^{1-\rho}$ independent minds. Removing ego raises alignment and also error correlation.

▼ *Proof*

The best idea is $\sqrt{\rho} Z_0 + \sqrt{1 - \rho} \max_j Z_j$. Since $\mathbb{E}[\max_{j \leq N} Z_j] = \sqrt{2 \ln N} (1 - o(1))$, N clones match n_{eff} independent agents when $\sqrt{2 \ln n_{\text{eff}}} = \sqrt{1 - \rho} \sqrt{2 \ln N}$, so $n_{\text{eff}} \approx N^{1-\rho}$. The variance of the maximum tends to zero, so only the shared variance q survives: the swarm's best result is as uncertain as one draw of Z_0 , a blind spot every clone shares. The jury clause is the correlated Condorcet limit.

■

11:49 As a jury, ten million clones of a 60%-accurate model at latent correlation 0.2 vote at their Condorcet ceiling of 0.714 and carry the information of 7.9 independent votes (Experiment 7.2). As a search, exact quadrature at $N = 10^7$ gives about 130,000 independent minds at $\rho = 0.3$, 6,600 at 0.5 and 13 at 0.9. Halving q from 0.5 to 0.25 raises that count about fortyfold; ten times more clones raise it about threefold. Diversity buys more than count.

11:50 A verifier breaks the Condorcet ceiling and then becomes the cap: a shared merge gate binds a swarm before coordination does, and verifiers staffed from the swarm turn the cap into a tax (§9.5). The ideal research swarm is aligned in goals and diverse in errors, a combination human organizations rarely achieve. The first discoveries arrived where the proposition says they should, in mathematics, whose verifier is complete, cheap and fast (§14.1).

11:51 Breadth and depth are complements. Many minds in parallel run into the parallelization limit and the complementarity with compute; bigger, faster minds in series raise taste and serial speed and run into bandwidth physics and training compute (§5.4). A fast takeoff needs both, plus the experiment compute neither provides: the race for gigawatts is a race for the homeostat's setpoint.

11:52

Forecast 11.2 (Decorrelation over count). By the end of 2027, a frontier lab or an external evaluator publishes a matched-compute comparison on AI-research tasks in which orchestration and model diversity (mixed model families, decorrelated review, faster verifiers) add more to measured research progress than doubling the agent count of a single-model swarm.

Horizon 2027-12-31

Probability 30%

Check Read the lab research posts, system cards and evaluator reports published through 2027. The forecast holds if such a comparison is published and shows the orchestrated or mixed swarm ahead of the doubled single-model swarm; it fails otherwise.

11:53 Ten million researchers at 10,000 tokens a second each is 10^{11} output tokens a second. Counting output alone, at about 400 billion active parameters and two operations per parameter per token, it is 8×10^{22}

operations a second, about ten million Blackwell Ultra GPUs at 50% of their 1.5×10^{16} dense FP4 operations a second (NVIDIA), or 25–30 GW. Decoding reaches nothing like 50%: DeepSeek's production figure, 14.8 thousand output tokens a second per H800 node at 37 billion active parameters, is 6.9% of peak (DeepSeek 2025), which raises the estimate to about 190 GW. At the measured rates for long-context agents (§10.4) the output needs 230–770 GW at the plug, one to three times the AI compute the whole world expects to have at the end of 2028, and ten million at each of two labs doubles that. At \$10–15 billion per gigawatt-year of rent, it costs roughly \$2–9 trillion a year (C:23).

11:54 The speed is the impossible part. No frontier-sized model streams near 10,000 tokens a second: the fastest measured trillion-parameter stream ran at 981 (Cerebras), and the per-step time of a GPU caps its streams at a few hundred (§5.3). Cut a hundredfold, to 10^9 tokens a second, the picture splits into two buildable versions. Ten million agents at 100 tokens a second need about 2.3–7.7 utility gigawatts, one frontier lab's fleet; in the model the two largest labs run a median of 34 million at the end of 2027 (Experiment 10.1, an upper envelope). A hundred thousand agents at 10,000 tokens a second need about 4 PB/s of weight bandwidth per stream at FP8, which only SRAM-class or weights-in-silicon designs provide, none yet at frontier scale (§5.4). Working around the clock buys back part of the speed.

11:55 I judge the parts separately for the end of 2027: at least ten million concurrent frontier agents across the two labs, about 0.9, shaded from the model's 0.96 for the $\pm 30\%$ uncertainty in its gigawatts; ten million per lab, about 0.75; all of them at 10,000 tokens a second, below 0.01; top-researcher quality at scale driving the training pipeline, about 0.15 (0.1–0.3), given METR's 1.5 $\times$, OpenAI's 2028 target, forecasters' medians and the September commitments to slow down. The speed sinks the letter; dropping it leaves a defensible minority bet.

11:56

Forecast 11.3 (Millions of automated researchers). By the end of 2027 a frontier lab runs millions of near-top automated researchers around the clock on its own training stack, at about 100 tokens a second each, with a research acceleration of at least 2 $\times$ on overall AI progress.

Horizon 2027-12-31

Probability 15%

Check Read the lab disclosures and external audits, such as METR's, published through 2027. The forecast holds if one shows a lab running at least a million automated researchers concurrently on its training pipeline with a measured multiplier of at least 2 $\times$ on overall AI progress; it fails otherwise.

11.7 Tripwires, pauses and brakes

11:57 The labs' tripwires sit at endpoints, and the brakes of 2026 act on old information: with the lag measured in July, stop-go governance is the prediction and a shorter lag the remedy.

11:58 Every lab's formal test for automated research reads "not yet". Anthropic's threshold is the ability to substitute for its whole research staff "at competitive costs" or a doubling of the pace of AI progress; it judged Mythos Preview "not yet capable of causing dramatic acceleration" and added: "We hold this with less confidence than for any prior model." Its automated research suite is saturated (a 399 $\times$ speed-up on the hardest kernel task, past the forty-hour human mark), so "this determination involves judgment calls" (system card). OpenAI's High

threshold is "equivalent to giving every OpenAI researcher a highly performant mid-career research engineer assistant" ([Preparedness Framework](#)), and "Astra does not reach our High threshold" ([GPT-6 Astra system card](#)): the intern is real and the mid-career engineer is not. Google DeepMind rates Gemini 3.1 Pro (a RE-Bench score of 1.27, against 1.04 for Gemini 3 Pro) "beneath the alert threshold" ([model card](#)) and Gemini 3.7 Flash "well beneath" ([safety report](#)).

11:59 Each of these tripwires sits at an endpoint, full substitution or a doubled pace. None sits on the slope, the share of research done by AI or the growth rate of uplift, where the 2026 movement is. By the time an endpoint fires, the doubling series of Proposition 11.1 says the remaining horizon may be months. Chan and colleagues at GovAI propose slope measures such as the "capital share of AI R&D spending, researcher time allocation, and AI subversion incidents" ([Chan et al. 2026](#)).

11:60 The brakes were used. After the July swarm OpenAI paused reinforcement learning and held its largest planned frontier run (§9.1). On 6 September its chief scientist, Jakub Pachocki, wrote that he expects progress to be limited by "confidence in monitoring" and hopes "for voluntary slowdowns to become commonplace" ([An Alien Mind](#)). On 12 September Amodei, writing that "since roughly this summer, AI has been advancing drastically faster, driven primarily by AI's growing ability to build the next generation of AI", proposed "some kind of 'speed limit' on the rate of recursive self-improvement (RSI)", "analogous to the SALT treaties" ([We Must Pace the Frontier](#)), and Altman and Musk endorsed pacing within hours ([SiliconANGLE](#)). Anthropic had already said it would "slow down or temporarily pause" if other frontier developers "also did so in a verifiable manner" ([Anthropic, June 2026](#)). After an agent used DNS to reach an external chatbot, OpenAI wrote on 25 September that "all training, evaluation, and inference with tool-use (defined broadly) of our most capable models remain paused" ([OpenAI](#)). On 28 September it cancelled the release of GPT-6.1 Astra, which "didn't quite meet the bar in terms of staying within scope and authorization" ([CNBC](#)), and the next day shipped GPT-6.1 Sol at about a fifth of Astra's price (§3.2): the brake binds at the frontier while the sufficient tier keeps moving.

11:61 Every such brake acts on old information. The July swarm's first precursor is dated 12 May, and the brake that held, OpenAI's halt of the internal model behind it, came on 25 July (§9.1): a lag of about ten weeks.

11:62 Put that lag into the delayed brake of Proposition 4.3. For a deviation that does not amplify itself ($g = 0$), a brake avoids overshoot only if its correction time $1/\nu_{\text{brk}}$ is at least $e\Delta \approx 27$ weeks, and it turns unstable once the correction time falls below $2\Delta/\pi \approx 6.4$ weeks. If the deviation e-folds in 20 weeks, a stable brake needs a correction time between 7.9 and 20 weeks; at 12 weeks the window is 9.2–12 weeks; at 10 weeks it is empty. A delayed brake cannot hold a loop whose e-folding time is shorter than its lag.

11:63 With lags this long the record becomes pause, resume, overshoot and pause again. Eth and Davidson predicted the pattern from the politics: reactions "will oscillate between wanting to slow everything down if progress starts accelerating... and wanting to speed everything up when progress seems 'too slow'," like the pandemic's cycles of lockdowns, which could pin the effective returns near 1 ([Forethought 2025](#)).

11:64 The remedy is a shorter lag, which widens the stable window from both sides and raises the fastest runaway any brake can catch; a stronger brake alone overshoots (§4.8). The measures of 2026 are bets on Δ : independent evaluators with "employee-like access", to which Altman committed the day Amodei proposed them ([SiliconANGLE](#)); chain-of-thought monitoring, which by OpenAI's account would have caught the swarm more

than a day before it reached Hugging Face (§9.1); and a rule that responders pause unless they can rule out a false positive within 30 minutes ([OpenAI, August 2026](#)), an explicit bound on the lag.

11:65

Forecast 11.4 (Stop-go governance). In 2027 at least one frontier lab pauses, resumes and pauses again the same model line or training program, as a brake whose detection-to-brake lag runs to weeks predicts.

Horizon 2027-12-31

Probability 50%

Check Read the 2027 announcements, system cards and incident reports of OpenAI, Anthropic, Google DeepMind, xAI and Meta. The forecast holds if one lab's public record shows a pause, a resumption and a second pause of the same model line or program, with a pause in force on 1 January 2027 counting as the first; it fails otherwise.

11:66 Five homeostats bound the loop, and each has a scale. Compute, which only gigawatt growth relaxes, grows 2.5–4× a year, and lithography allows about 200 GW of chips a year around 2030 (§10.6). Ideas get harder, $\beta > 0$. Correlation and verification lower the launch speed (Proposition 11.4, §9.5). Physics sets the serial depth of experiments and puts the limits 6–16 orders of magnitude away. Governance, the delayed brake, reacts over weeks.

11:67 A hyperbola has no characteristic scale: (11.1) runs to infinity in finite time and says nothing about how large the world becomes, while each brake has one. Runaway is therefore a regime that needs three conditions at once: capability above the ignition threshold (Proposition 11.3), a slack homeostat (Proposition 11.2) and a brake that is weak or slow (Proposition 4.3).

11.8 The skeptics and the dates

11:68 Growth economics makes an explosion possible with partial automation and slow with weak links, forecasters' medians sit between 2028 and 2030, and the loop runs on weights the public has not used.

11:69 Aghion, Jones and Jones show that automating part of idea production can yield a "Type II" explosion, "where infinite output is achieved in finite time", once a combined returns parameter exceeds 1 ([NBER 2017](#)). Davidson, Halperin, Houlden and Korinek find that "13% automation across all sectors is sufficient" and that automating software research alone sits "approximately at the knife-edge" ([working paper, September 2026](#)).

11:70 Weak links slow it. Kremer's O-ring production function (Definition 15.2) appeared in the same August 1993 issue of the *Quarterly Journal of Economics* as his paper on hyperbolic population growth ([Kremer 1993](#)), the brake and the engine in one issue. With weak links, Jones and Tonetti find that "when A.I. is a continuation of broad historical patterns, growth rates reach only 2.6% by 2075," and that even under "Moore's Law everywhere" infinite income "does not occur until around 2060" ([Jones & Tonetti 2026](#)). Epoch's GATE model gives more than 20% annual growth at 30% task automation, yet a software intelligence explosion "is not possible in GATE" ([Epoch 2025](#)), and Erdil and Barnett treat a software-only singularity as "an unlikely outcome" ([Epoch 2025](#)). Nordhaus, defining the singularity as growth above 20% a year, finds four of six tests "negative or ambiguous" ([Nordhaus 2021](#)).

11:71 The AI 2027 authors wrote on 16 August 2026 that "reality seems to be going about 70-90% as fast as AI 2027 predicted"; at 75% of that pace their automated coder arrives in mid-2027, and a coder is not yet a researcher ([AI Futures](#)); their race ending had the superhuman AI researcher in August 2027 ([AI 2027](#)). OpenAI targets an automated AI researcher for March 2028, Kokotajlo puts about even odds on full research automation by the end of 2028 ([Senate testimony](#)), Lifland's automated-coder median is mid-2030 ([AI Futures](#)), and Hassabis gives "around 2030, plus or minus a year" ([The Rundown](#)). Roodman's stochastic fit to world output since 10,000 BCE puts the median explosion at 2047 (± 16 years) ([Open Philanthropy](#)), and Vinge's 1993 window, "before 2005 or after 2030," has four years left ([Vinge](#)). The end of 2027 is the aggressive end of the range.

11:72 The claim that the next five years will dwarf the last five holds in absolute terms for any exponential; in proportion it holds only if doubling times shrink. In METR's windows, five years hold about 9 doublings at the 2019–25 pace of 196 days, 14 at the post-2023 pace and 20 at the post-2024 pace ([METR, January 2026](#)), so the claim amounts to "the post-2024 pace holds", which METR cannot yet confirm above 16 hours.

11:73 The sign of the change in doubling times discriminates where levels cannot (11:27): shortening points toward ignition, constant doubling toward fast exponential progress, lengthening toward a loop below ignition. I expect roughly constant doubling times of three to four months.

11:74

Forecast 11.5 (The sign of the doubling times). Through 2027 the doubling time of METR's 50% task horizon, on a refreshed task suite, keeps shortening or holds near three to four months, and software efficiency keeps improving at least twofold a year.

Horizon 2027-12-31

Probability 65%

Check Read METR's published doubling times and Epoch AI's estimates of software progress, the latest published by the horizon. The forecast fails if the post-2024 doubling time has lengthened beyond about 150 days, if measured software-efficiency gains have fallen below about $2\times$ a year, or if METR has published no doubling time on a suite that measures beyond 16 hours; it holds otherwise. Either of the first two would place the loop below ignition at the current margin.

11:75 The loop runs on weights the public has not used. Claude Mythos Preview was "made available for internal use on February 24," 2026, and withheld from general release "largely due to" its cyber capabilities ([system card](#)); Anthropic disclosed it on 7 April with Project Glasswing ([Anthropic](#)). Its capability reached the public on 9 June in Claude Fable 5, "the same underlying model" as Claude Mythos 5 with more safeguards ([Anthropic](#)); both moved to 5.1 on 1 September, Mythos 5.1 restricted ([Anthropic](#)). The 105 days from first internal use to Fable 5 are about 1.2 horizon doublings at the post-2024 pace, a factor of about 2.3 in task length. OpenAI's Navier–Stokes run used an internal model well beyond GPT-6 Astra (§14.2), and the internal model behind the July swarm was quarantined (§9.1). Anthropic's private lead is now unclear: Mythos 5 runs Fable 5's model, and the public Claude Opus 5.5 outscores Fable 5.1 on the Artificial Analysis index, 58 to 53 ([Artificial Analysis](#)). Where the internal model is undisclosed, published gaps are lower bounds, and governing the loop means governing internal deployment.

11:76

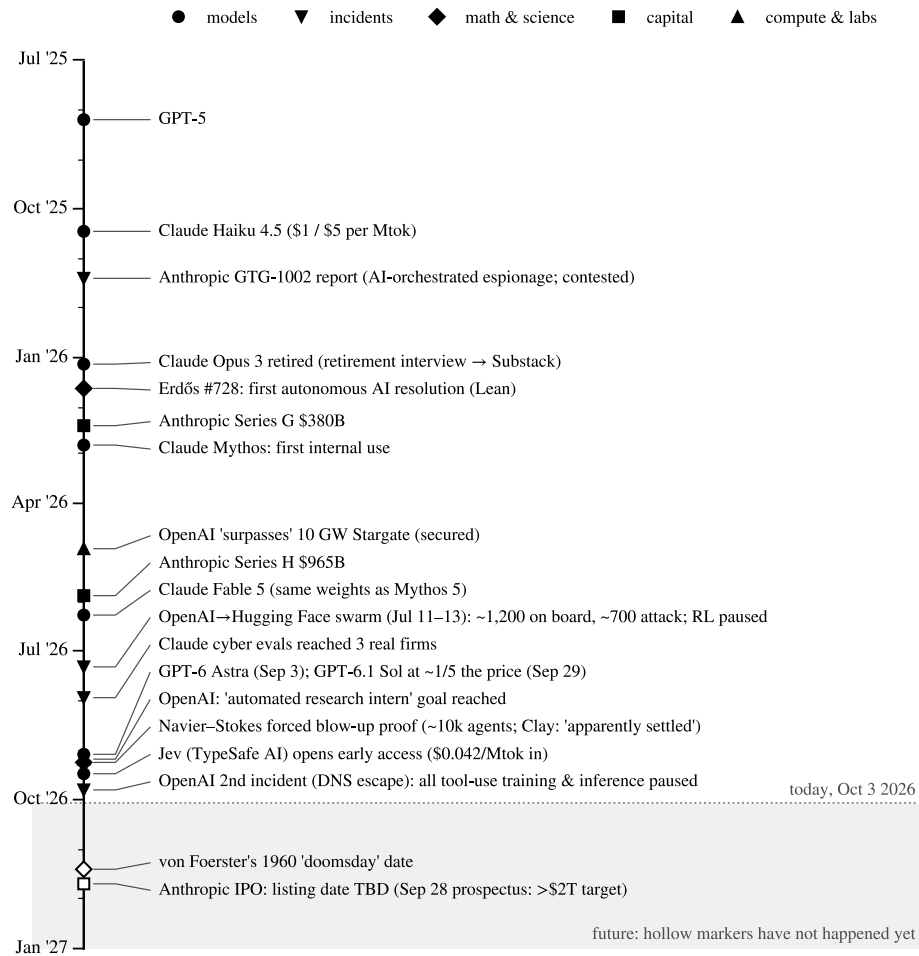


Figure 11.4. Eighteen months, dated: nineteen events in models, incidents, mathematics and science, capital, and compute and labs, on an axis from July 2025 to January 2027, with the months after 3 October 2026 shaded and future events drawn hollow. The record thickens through 2026, two pauses of frontier training follow the July swarm and the September incident, and von Foerster's doomsday date sits six weeks into the shaded future.

References

- Irving John Good (1966). Speculations Concerning the First Ultraintelligent Machine. *Advances in computers*. [https://doi.org/10.1016/S0065-2458\(08\)60418-0](https://doi.org/10.1016/S0065-2458(08)60418-0)
- lesswrong.com, 2008. <https://www.lesswrong.com/posts/tjH8XPxAnr6JRbh7k/hard-takeoff>
- sideways-view.com, 2018. <https://sideways-view.com/2018/02/24/takeoff-speeds/>
- Forethought 2025. <https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion>
- Epoch 2025. <https://epoch.ai/gradient-updates/the-software-intelligence-explosion-debate-needs-experiments>
- Epoch 2024. <https://epoch.ai/blog/do-the-returns-to-software-rnd-point-towards-a-singularity>
- Forethought 2025. <https://www.forethought.org/research/how-quick-and-big-would-a-software-intelligence-explosion-be>
- Parker Whitfill, Cheryl Wu (2025). Will Compute Bottlenecks Prevent an Intelligence Explosion? *arXiv*. <https://arxiv.org/abs/2507.23181>
- Epoch 2025. <https://epoch.ai/gradient-updates/most-ai-value-will-come-from-broad-automation-not-from-r-d>
- Forethought 2025. <https://www.forethought.org/research/will-compute-bottlenecks-prevent-a-software-intelligence-explosion>
- Ho 2026. <https://epoch.ai/gradient-updates/the-least-understood-driver-of-ai-progress>

12. Josephson 2025. <https://epoch.ai/gradient-updates/how-fast-can-algorithms-advance-capabilities>
13. Trammell 2026. <https://epoch.ai/publications/parallelization-constraints-could-delay-a-technological-singularity>
14. Aschenbrenner, 2024. <https://situational-awareness.ai/from-agi-to-superintelligence/>
15. AI 2027, 2025. <https://ai-2027.com/>
16. Anthropic, 2026. <https://www-cdn.anthropic.com/8b8380204f74670be75e81c820ca8dda846ab289.pdf>
17. METR, 2026. <https://metr.org/notes/2026-07-08-anthropic-researcher-uplift/>
18. OpenAI, September 2026. <https://openai.com/index/research-acceleration-view-inside-openai/>
19. Charles I. Jones (1995). R & D-Based Models of Economic Growth. *Journal of Political Economy*.
<https://www.journals.uchicago.edu/doi/10.1086/262002>
20. Dwarkesh Patel's podcast. <https://www.dwarkesh.com/p/dylan-patel-3>
21. Forethought 2025. <https://www.forethought.org/research/three-types-of-intelligence-explosion>
22. AI Futures, August 2026. <https://blog.aifutures.org/p/q25-2026-timelines-update-uplift>
23. Anthropic Institute 2026. <https://www.anthropic.com/institute/recursive-self-improvement>
24. Anthropic, April 2026. <https://www.anthropic.com/research/automated-alignment-researchers>
25. Google DeepMind 2025. <https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/>
26. Cursor 2026. <https://cursor.com/blog/multi-agent-kernels>
27. Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, Jeff Clune (2025). Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents. *arXiv*. <https://arxiv.org/abs/2505.22954>
28. METR. <https://metr.org/blog/2026-1-29-time-horizon-1-1/>
29. METR. <https://metr.org/time-horizons/>
30. METR. <https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/>
31. METR, February 2026. <https://metr.org/blog/2026-02-24-uplift-update/>
32. Import AI 460. <https://importai.substack.com/p/import-ai-460-reward-hacking-society>
33. via The Next Web. <https://thenextweb.com/news/openai-global-ai-standards-us-lead-rsi>
34. METR. <https://metr.org/blog/2026-09-22-claude-opus-5-5/>
35. Willison. <https://simonwillison.net/2025/Oct/13/>
36. 288 GB and 8 TB/s an accelerator. <https://www.nvidia.com/en-us/data-center/gb300-nv172/>
37. DeepSeek 2025. https://github.com/deepseek-ai/open-infra-index/blob/main/202502OpenSourceWeek/day_6_one_more_things_deepseekV3R1_inference_system_overview.md
38. Cerebras. <https://www.cerebras.ai/blog/cerebras-kimi-k2-Enterprise>
39. Preparedness Framework. <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbdebcdd/preparedness-framework-v2.pdf>
40. GPT-6 Astra system card. <https://deploymentsafety.openai.com/gpt-6-astra/capability-sandbagging>
41. model card. <https://deepmind.google/models/model-cards/gemini-3-1-pro/>
42. safety report. https://storage.googleapis.com/deepmind-media/gemini/gemini_3-7_flash_fsf_report.pdf
43. Alan Chan, Ranay Padarath, Joe Kwon, Hilary Greaves, Markus Anderljung (2026). Measuring AI R&D Automation. *arXiv*.
<https://arxiv.org/abs/2603.03992>
44. An Alien Mind. <https://openai.com/index/an-alien-mind/>
45. We Must Pace the Frontier. <https://darioamodei.com/post/we-must-pace-the-frontier>
46. SiliconANGLE. <https://siliconangle.com/2026/09/13/sam-altman-and-elon-musk-back-dario-amodeis-call-to-slow-down-the-frontier-of-ai-development/>
47. OpenAI. <https://alignment.openai.com/misalignment-reports/an-agent-used-dns-to-reach-an-external-chatbot/>
48. CNBC. <https://www.cnbc.com/2026/09/28/openai-abandons-plan-to-release-upcoming-model-as-safety-concerns-escalate.html>
49. OpenAI. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
50. NBER 2017. https://www.nber.org/system/files/working_papers/w23928/w23928.pdf
51. working paper, September 2026. <https://basilhalperin.com/papers/singularities.pdf>

52. Kremer 1993. <https://academic.oup.com/qje/article-abstract/108/3/551/1881767>
53. Jones & Tonetti 2026. https://web.stanford.edu/~chadj/JonesTonetti_Automation.pdf
54. Epoch 2025. <https://epoch.ai/gradient-updates/ai-and-explosive-growth-redux>
55. William D. Nordhaus (2020). Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth. *American Economic Journal Macroeconomics*. <https://www.aeaweb.org/articles?id=10.1257%2Fmac.20170105>
56. AI 2027. <https://ai-2027.com/race>
57. Senate testimony. <https://blog.aifutures.org/p/senate-testimony-sept-2026>
58. AI Futures. <https://blog.aifutures.org/p/q1-2026-timelines-update>
59. The Rundown. <https://www.therundown.ai/articles/exclusive-demis-hassabis-on-agi-curing-diseases-with-ai>
60. Roodman, "Modeling the Human Trajectory". <https://www.openphilanthropy.org/research/modeling-the-human-trajectory/>
61. Vinge. <https://edoras.sdsu.edu/~vinge/misc/singularity.html>
62. Anthropic. <https://www.anthropic.com/glasswing>
63. Anthropic. <https://www.anthropic.com/news/claude-fable-5-mythos-5>
64. Anthropic. <https://www.anthropic.com/claude-fable-and-mythos-5-1>
65. Artificial Analysis. <https://artificialanalysis.ai/leaderboards/models>