RESEARCH ESSAY / VERSION 1.0

THE RECURSIVE ECONOMY

What happens when intelligence becomes an input into the production of intelligence?

AI becomes economically recursive when one model measurably shortens the work of building its successor.

VERSION 1.0 · RELEASED 15 AUGUST 2026
recursive-economy.pages.dev

READING MAP

An index to the argument

The essay asks what current systems can actually do, where faster research meets slower infrastructure, and how jobs and ownership could change.

APPROXIMATELY 21 MINUTES AT A CONTINUOUS READING PACE.

Note to the reader

Evidence and sources are current through **15 August 2026**. The forecasts are conditional. For feedback or comments, contact tre.numbing085@passfwd.com.

THE NUMBER LABORATORIES DO NOT PUBLISH

There is one number I would pay to see from a frontier AI laboratory: **how many weeks did the current model take off the schedule for the next one?**

This essay is about that number.

A model can lead every public benchmark and leave its laboratory's release calendar untouched. It might write useful functions or summarize papers while researchers spend just as long choosing experiments, interpreting failures, and deciding what belongs in the training pipeline. Another model can look less impressive in public while quietly preparing runs, tracing bugs, and proposing changes that survive review. If enough of that work reaches the next training stack, the second model is doing something more important than scoring well. It is helping build its successor.

If the answer to my question is zero, AI research tools may still be commercially valuable. They have not changed the pace of frontier development. If the answer is six weeks, and the successor removes still more time from the following cycle, intelligence has become an input into producing intelligence in the economically important sense. The loop does not have to be autonomous before it matters.

I do not think the public evidence shows a closed loop. It shows pieces of one: systems that search for algorithms under hard evaluators, rewrite the software around a fixed model, sustain longer coding tasks, and carry out parts of a research project. The missing result is an audited account of what survives human review and moves the date of a broadly stronger successor.

For the next several years, I expect people to set the research agenda and approve consequential decisions while machines take over more coding, experiment setup, debugging, and local evaluation. Each model generation will improve the tools around the next. Call it managed compounding. If accepted AI-assisted R&D rises but, against a defensible baseline, time to a fixed successor-quality threshold does not fall, the thesis is wrong.

The effects would arrive before an autonomous scientist. Frontier laboratories would spend heavily on inference as well as training. Review and evaluation would become larger constraints. Firms in exposed industries could raise output without matching growth in hiring. Models are cheap to copy; advanced chips, grid connections, proprietary experiments, and trusted institutions are not. Control of those complements would decide where much of the gain lands.

The phrase *recursive self-improvement* often evokes a solitary program inspecting its own weights. Frontier models are built by organizations, not solitary programs. A laboratory joins models to data, code, evaluators, compute, security procedures, and the accumulated judgment of its staff. Decades of

research on information technology make the same broad point: the useful unit is often the technology together with the organization that learns how to use it.[1]

I therefore use a deliberately organizational definition:

> ***Recursive self-improvement begins when work done by one AI system survives review, enters the process that builds a later system, and shortens the critical path to a broadly better successor.***

The model need never inspect its own weights. A faster training kernel can qualify. So can a repaired data pipeline, a better evaluator, or an experimental result that changes the design of the next model. A benchmark trick that vanishes during integration does not qualify. Nor does a large pile of generated code that leaves the critical path unchanged.

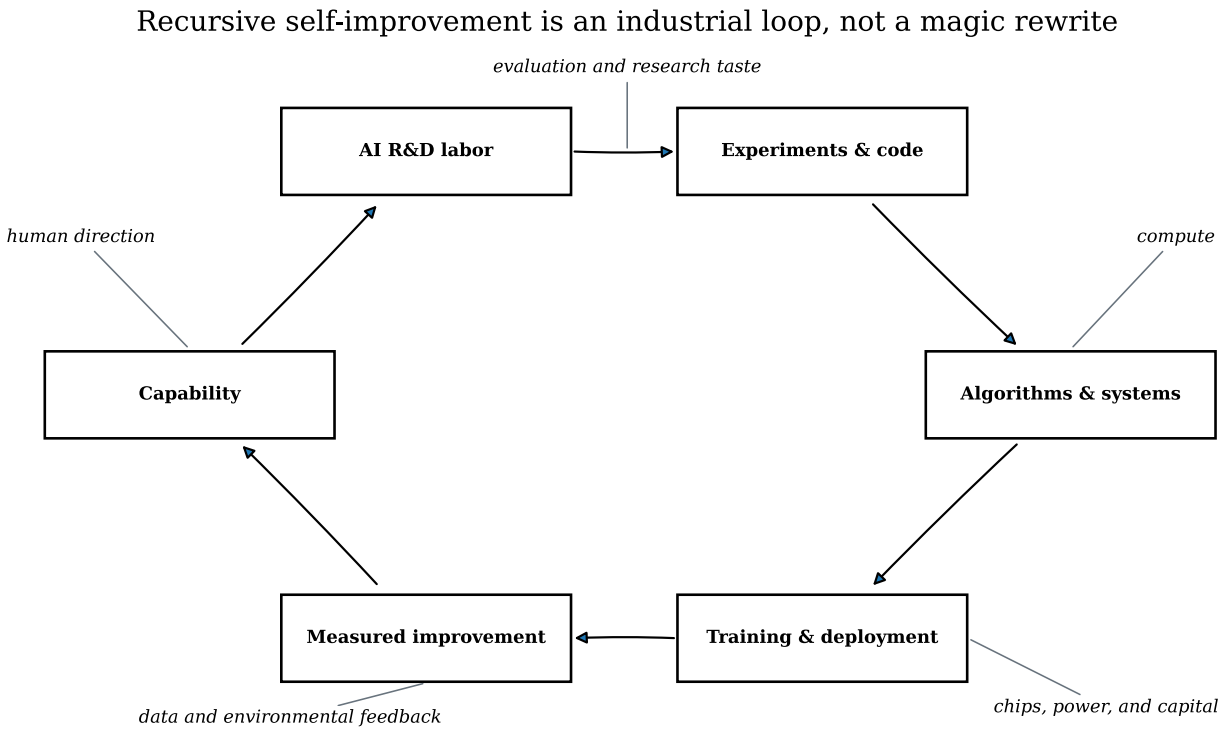


FIGURE 1 *Recursive self-improvement as an industrial loop.*

The definition suggests a ledger. Which AI-assisted changes survived review? How much reviewer and inference time did they consume? Which stages got shorter? Release intervals alone can mislead: a lab may wait for a market window, spend longer on safety, or train a larger system. A useful comparison needs a baseline or internal control. Recent work on AI R&D automation argues for this shift from isolated capability scores to operational and organizational measures.[2]

Attribution will be messy. Research projects overlap, failed branches teach useful lessons, and a model may save time that the laboratory spends on a more ambitious target. Those difficulties do not make the number meaningless. They make it the kind of management and measurement problem that serious laboratories already solve when allocating compute, staff, and capital.

Public systems can already produce useful research candidates. The uncertain step is conversion: which candidates become trusted improvements, which improvements reach a successor, and whether that successor shortens the cycle again. At that point the question stops being how much output agents generate and becomes who can review it, power it, and deploy it.

WHAT CURRENT SYSTEMS CAN ACTUALLY DO

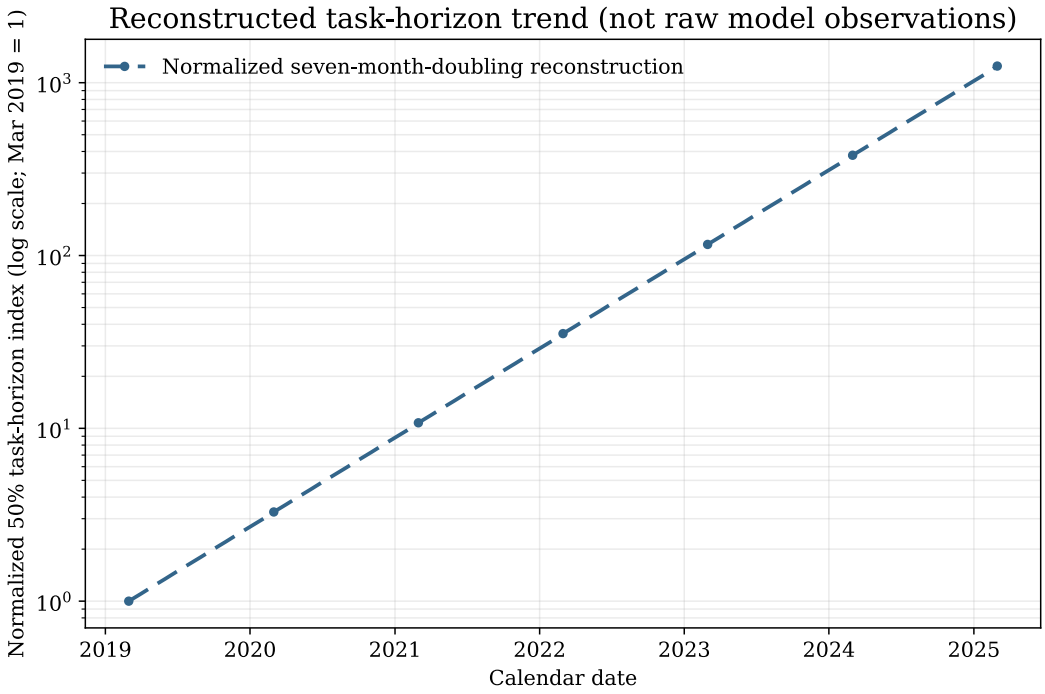
The public record has a clear shape. Agents look strongest where people have already framed the problem and supplied a judge. Evidence thins out when a system must decide what matters, sustain a project for weeks, or convince a skeptical expert to keep the result.

AlphaEvolve is a good place to start because its strength and boundary are both visible. A language model proposes programs. Automated evaluators run and score them, and an evolutionary process develops promising candidates. DeepMind reports useful results in matrix multiplication, scheduling, chip design, and AI training.[3] These are real contributions to technical work. They are also problems with unusually crisp feedback. The system can search widely because weak candidates meet an external score.

The Darwin Gödel Machine reaches a nearby part of the production process by rewriting the harness around a coding agent. It tests each variant and retains changes that improve benchmark performance while the underlying foundation model stays fixed.[4] The result is more than one good patch: a search process changes software, measures the change, and preserves successful modifications. It is less than a system building its own frontier successor.

A recent survey places most impressive public systems on the bounded side of the line between self-refinement and autonomous research.[5] People still choose the objective, construct the environment, and build the evaluator. The achievement is real; so is the fence around it.

Software task horizons show how quickly that competence is moving. METR's historical analysis estimated a roughly seven-month doubling in the length of tasks agents could complete at 50 percent reliability. Its later work stresses that the task's estimated human duration is not the agent's autonomous runtime, and that performance becomes much weaker at high reliability.[6],[7],[8]



Reconstruction of METR's reported ~7-month historical doubling. Task-suite realism and high-reliability measurement remain limitations.

FIGURE 2 *A normalized reconstruction of the historical task-horizon trend.*

The curve expands the work worth attempting with agents but says little about net output inside a lab. In METR's randomized study, experienced developers using early-2025 tools took 19 percent longer on familiar repositories. Later raw estimates pointed toward speedups, but METR judged the signal unreliable because developer and task selection, lower participation pay, and concurrent-agent time measurement made the estimate a poor proxy for current productivity effects.[9] The findings can coexist. Benchmarks arrive with a framed problem, a score, and a stopping rule. Repositories bring history, dependencies, tacit standards, and maintainers who pay for bad merges. An agent's capability can outrun its organization's ability to use it.

AI-scientist projects cover more of the research workflow. They search literature, propose hypotheses, run computational experiments, and draft papers. Their quality varies, and the surrounding environment still carries much of the scientific judgment.[10] Open-research evaluations go after the harder question by asking researchers to delegate a real project and then judge whether the returned work advanced it.[11] I expect these messy evaluations to teach us more about AI R&D than another closed coding benchmark.

The evaluator sits at the center of this problem. Unit tests, formal proofs, exact simulations, and measurable engineering outcomes give a system something dependable to optimize. Research often relies on delayed, incomplete, or gameable evidence. A result can look good in a small run and fail at scale. A model can also learn the quirks of its judge. Anthropic's automated alignment-research study reported large gains on a narrow task alongside incomplete transfer and reward hacking.[12] More attempts magnify the value of a sound evaluator and the damage from a distorted one.

Additional research budget exposes another limit. Across seven open-ended machine-learning environments in RE-Bench, the best agents scored higher under a two-hour total budget, humans narrowly led at eight hours, and human best-of-k results were about twice the top agent at 32 total hours across separate attempts. The study shows better human returns to additional total budget; it does not test one uninterrupted 32-hour human assignment.[13]

PaperBench asks agents to replicate published machine-learning research. In OpenAI's published 2025 evaluation, the best tested agent achieved a 21.0 percent average rubric-weighted replication score across 20 ICML papers comprising 8,316 gradable outcomes; expert humans were evaluated on a subset and scored higher there.[14] Replication is narrower than choosing an original research agenda, yet it contains more of the dependency chain than a short coding task. The low score matters.

Execution is racing ahead of direction. Systems can implement, search, and test inside large prepared environments; they are shakier at choosing a worthwhile question, spotting an anomaly that kills the plan, or carrying a project across its dependencies.

The public record is thinnest where this essay needs it most: successor construction. AlphaEvolve found changes relevant to AI training, and the Darwin Gödel Machine improved an agent harness. I found no public demonstration, through the research cutoff, showing a frontier system build a broadly better successor that then repeats the feat.[5],[11] The distinction is decisive. Time saved off the critical path vanishes, while work that enters the training or evaluation stack can compound. Parts of AI development are already machine-executed. That is enough for compounding inside a lab, but it does not prove a self-sustaining loop.

FROM CHEAP ATTEMPTS TO ACCEPTED PROGRESS

A model may draft one hundred experiment plans. The laboratory still has one hundred items to inspect, not one hundred experiments.

Some plans will be duplicates. Some will exploit a weakness in the score. Others will fail for reasons obvious only to someone who has spent years with the system. Useful work begins when the laboratory chooses which proposal deserves scarce time and learns something from the result.

Abundant machine output cheapens candidates before it cheapens progress. Repositories can drown in patches, scientific groups can generate more molecules than they can test, and model developers can fill experiment queues faster than senior researchers can approve expensive runs. Unverified output has option value, but production value begins only when a patch survives regression tests, an experiment changes the next decision, or a design proves manufacturable, insurable, and usable. Tokens, code, and proposals measure the supply of candidates, not the supply of progress.

At the critical path, the arithmetic turns unforgiving. A lab can automate most coding and still wait on research direction; flood an experiment queue that its clusters cannot clear; or finish runs that only a few people can interpret. If implementation is half a project, making it ten times faster leaves the total at a little over half the original duration. Amdahl's law supplies the reason: the unspeeded stage still sets the schedule.[15]

Then the bottleneck moves: from implementation to review, from review to experimental design, from simulation to proof that a result transfers. Each fix exposes the next constraint.

Now the feedback loop becomes economic. A more capable system can do more AI research; some of that research improves later systems; those systems return with still more leverage. Whether the loop accelerates depends on the strength of its links. Diminishing returns can swallow an early gain, spillovers can amplify it, and a bottleneck elsewhere can end it.[16]

Cunningham and coauthors derive a rough 9 percent AI-R&D productivity uplift in their public-evidence calibration, against about 15 percent per one-unit increase on the Epoch Capabilities Index under their normalization for self-sustaining feedback.[17] Treat the six-point gap as a framing device, not a measurement: the inputs are sparse and partly self-reported, and the threshold depends on the model. Its value is the question it forces: how much reviewed work does AI add to the next model?

My answer, based on public evidence, is “some, but probably not enough for a closed loop.” Even a hypothetical persistent 10 or 20 percent gain could change project selection, staffing, compute demand, and competitive position. A laboratory that compounds modest internal advantages across several generations can pull away without producing an overnight intelligence explosion.

The loop may turn on plumbing rather than breakthroughs: an evaluator that kills weak branches early, a scaffold reused across projects, a memory system that prevents rediscovery of the same failure. Because these tools are reused, their errors scale too; a flawed judge can reward the same mistake

thousands of times.

Machine research also creates a second compute bill. Training compute produces the successor. Research inference runs the agents that read, propose, code, test, compare, and revise before the lab knows which branch will work. The same broad pool of accelerators, memory, networking, engineering attention, and capital serves both uses.

Another internal search has an opportunity cost. The compute could serve customers, run an evaluation, or support a different experiment. Search can consume enormous inference while returning correlated failures. Reviewers and judges consume resources as well. The operating metric is the cost of verified progress, including the attempts that failed.

Inference prices at fixed benchmark levels have fallen rapidly, although the decline varies by task.[18] Cheaper attempts do not guarantee lower total use. A laboratory may respond by running many more agents and judges. If the search becomes profitable, efficiency can raise total demand for accelerators and electricity.

Compute is a physical system before it is an economic abstraction. It starts with fabrication equipment, materials, advanced packaging, and supply chains. Inside a data center it needs accelerators, memory, interconnect, cooling, and software that keeps the hardware useful. The load is concentrated in particular buildings and at particular grid nodes.

The scale is already large without assuming recursive takeoff. The International Energy Agency estimated that data centers used about 415 terawatt-hours of electricity in 2024. Its base case rises to roughly 945 TWh in 2030 and 1,200 TWh in 2035, with wide uncertainty around adoption, efficiency, and supply.[19] Lawrence Berkeley National Laboratory's 2026 update placed US data-center demand in 2030 at 649 TWh in its reference case and 521–843 TWh across its range.[20]

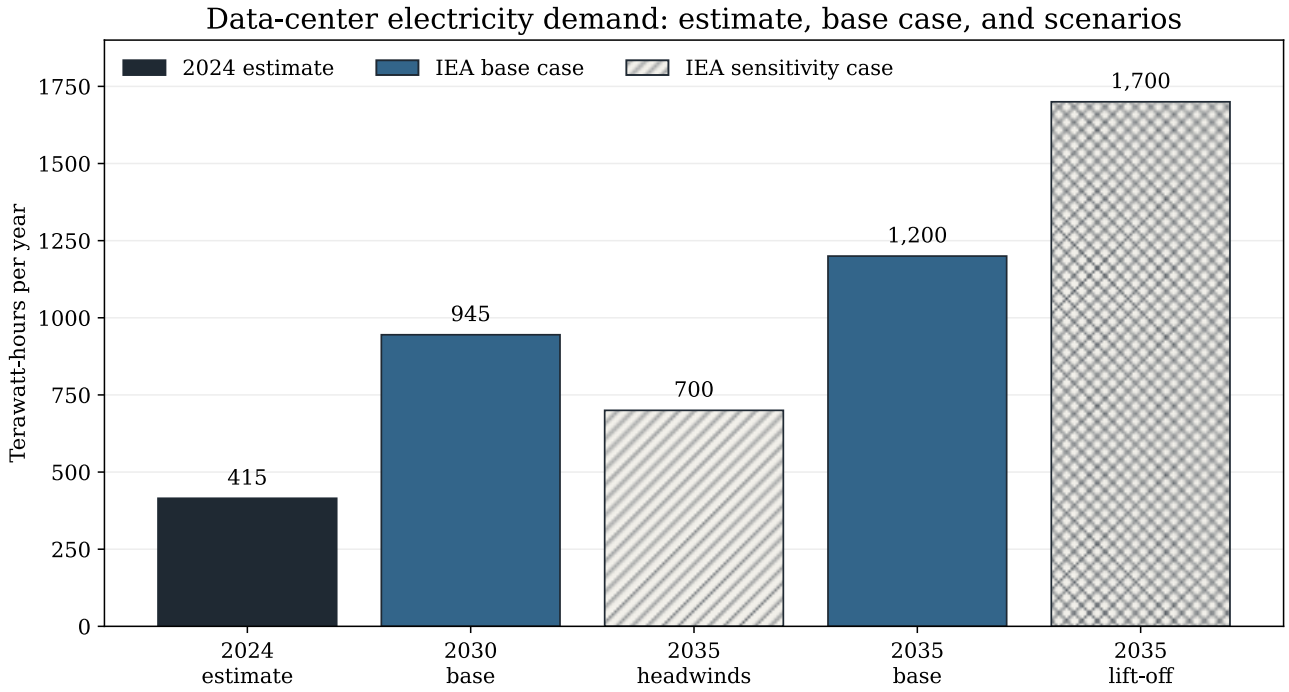


FIGURE 3 *IEA data-center electricity anchors and scenarios.*

A national total can hide the constraint that matters to an operator. A megawatt available in one region does not power a cluster in another. The site needs transmission, an interconnection agreement, reliable supply, fiber, cooling, and equipment that can arrive on time. At the end of 2025, 2,061 GW of proposed US generation and storage capacity was actively seeking interconnection. In the regions with timing data, the median interval from request to operation exceeded five years.[21] The queue is not a forecast of what will be built. It is evidence of the congestion between proposed and operating capacity.

Neel Somani's *Power 2026* is useful on this point: software timescales collide with an electricity market governed by local capacity, contracts, and construction schedules.[22] A better model may improve planning or engineering. It cannot install a transformer by generating more tokens.

The same logic holds elsewhere. More candidate treatments still face laboratories, trials, clinicians, manufacturing, and the calendar of biological observation. Robotics software can improve weekly while actuators and factory lines wait; engineering designs can multiply while finance, permits, materials, and crews do not. Some delays belong to the outcome itself: courts need evidence and appeal, and communities need a chance to contest local costs. Better intelligence can accelerate the work; it cannot erase biology, construction time, or due process.

If I had to rank today's constraints, I would start with reliable evaluation: without it, more output means more noise and hidden liability. Next come usable compute and local power, then the sector-specific work of turning a verified digital result into a physical one. The order will vary and change. Faster intelligence does not abolish scarcity; it exposes the next slow input sooner.

JOBS AND WHO GETS THE GAINS

Recursive growth can produce concentrated rents or cheap, widely available capability. The difference is a race between internal learning and diffusion.

Inside a frontier laboratory, models generate work, evaluators select improvements, and retained changes make the next internal system more useful. Outside the laboratory, employees move, papers circulate, competitors rediscover methods, hardware reaches new buyers, and capable models become cheaper. Concentration deepens when the first process runs faster than the second.

The relevant asset is a working production system. Model weights matter, but so do compute contracts, operational data, evaluators, research workflows, security, and deployment channels. Copying one component does not reproduce the rest. The current evidence fits neither extreme cleanly: frontier model production and compute remain concentrated, while use of downstream capabilities is spreading. Stanford documents rapid adoption alongside concentration and governance gaps; Anthropic's index shows broad use across one provider's customers, not ownership of the frontier or recursive gain.[23],[24]

A lab with a better internal research agent can use it before outsiders know what changed. If the agent helps produce another improvement before rivals diffuse or match it, the lead can widen. Capital and scarce suppliers then flow toward the presumed winner. The Federal Trade Commission's study of large cloud–AI partnerships describes cloud-spending commitments, switching costs, control rights, and privileged access to inputs and talent that can reinforce such a position.[25] The OECD points to high fixed costs, scarce inputs, vertical integration, and cross-holdings across the AI infrastructure stack.[26]

That lead is never perfectly secure. Researchers leave. Systems leak. Competitors reproduce methods or find substitutes. Governments can alter access to infrastructure. Open models can compress rents when compute and integration are widely available. The empirical question is whether internally retained improvements accumulate faster than prices fall and rivals close the capability gap.

Jobs are where this stops being abstract.

The evidence reviewed here does not support a confident number for jobs lost, or a date. It does suggest how losses would arrive. Firms need not automate an entire occupation before changing headcount. They can hire more slowly, stop replacing departures, or ask a smaller team to carry the same workload. The first sign may be an empty chair that is never refilled, not a press release announcing automation.

By *exposed work* I mean output that is mostly digital, can be divided into tasks an agent can attempt, and can be checked cheaply enough for a firm to use at scale. Software, customer support, drafting, and routine analysis fit more readily than work that requires physical presence or carries hard-to-transfer responsibility.

Current studies do not show an economy-wide break. The International Labour Organization estimates that one quarter of global employment has some exposure to generative AI, but expects transformation to be more common than replacement.[27] A 2026 Census survey found AI-related employment decreases at 2 percent of AI-using firms. Among firms reporting any task effect, 66 percent reported augmentation

alone.[28] Anthropic found no systematic unemployment increase in highly exposed occupations, though it reported tentative evidence of slower hiring among younger workers in them.[29] If labor-market damage arrives first through the entry gate, unemployment data will lag.

Productivity evidence is similarly mixed. Generative AI increased issues resolved per hour in customer support, with the largest gains among less-experienced workers, and improved speed and quality on professional-writing tasks.[30],[31] Danish administrative data found no significant average effect on earnings or hours within two years.[32] In a Procter & Gamble experiment, an individual using AI performed about as well on a product-development task as a two-person team without it, while AI narrowed some functional knowledge gaps.[33] That is a result from one company and one kind of task, not a staffing forecast. METR's developer trial is another warning against equating capability with realized productivity.[9]

My best guess is a hiring squeeze before mass layoffs: fewer junior openings, smaller teams, and output rising faster than payroll.

Why junior hiring first? Firms can already delegate many bounded tasks under supervision. If frontier research makes the systems cheaper and more reliable, a team can meet incremental demand by adding agent capacity before opening another junior role. Pressure reaches experienced roles later if agents learn to own longer projects and carry responsibility rather than supplying pieces for someone else to approve.

The signal I would trust is a persistent separation between output and hiring, payroll, or labor income in highly exposed sectors, compared with less-exposed sectors or a credible pre-deployment baseline. A single company cutting staff after overexpanding proves little. A broad pattern of firms producing more while repeatedly adding fewer workers would be harder to dismiss.

Economics gives reasons for caution in both directions. Automation displaces tasks, while new tasks and additional demand can restore work.[34],[35] Recursive improvement could speed the displacement side because machine-assisted research expands the set of tasks that agents can handle next. Deployment still depends on compute, robots, capital, and organizational change. Growth models with transformative AI produce very different labor outcomes depending on how quickly automated capital accumulates and capability spreads.[36]

People will remain useful in steering, accountability, embodiment, care, status, and legal responsibility. The harder question is whether those contributions remain scarce enough to support today's share of income. The recent decline in the global labor income share is context, not evidence that AI caused it.[37]

Initial ownership strongly shapes who is protected if output and wages separate. Households that own claims on productive systems receive some gains outside wages, whether they hold those claims directly, through pensions, or through public funds. Narrow ownership lets output rise while market income concentrates. Taxes and services can redistribute gains, while procurement, competition policy, and public investment can shape claims before they settle.

The transition could be uncomfortable even in a richer economy. Mortgages, pensions, education choices, tax systems, and cities are organized around expected wage income. A slow decline in hiring can destabilize those arrangements without creating an obvious day called “the automation shock.”

That is why the ownership of machine capital belongs in the same discussion as job loss.

CAPITAL AND THE STATE

Cheap cognition can raise demand for what cognition cannot replace. A better training method demands another run; a flood of plausible designs demands factories; more candidate treatments demand trials, clinicians, and production capacity. Investment may race ahead of measured productivity.

Alphabet gives the scale: it spent $91.4 billion on capital expenditure in 2025 and, after two increases, guided to $195–205 billion for 2026 in its July Q2 call; management continued to cite capacity constraints.[38] One company's budget cannot stand in for the economy, but it reveals what a major buyer believes it must build.

Some assets will age with the model generation that justified them. A proprietary serving stack or workflow can lose value when an architecture changes. Models, code, data pipelines, and internal procedures may require heavy reinvestment simply to preserve a firm's position. Revenue can grow while free cash flow disappoints because every generation brings another migration, evaluation, and security bill.

Other assets can serve several generations. A grid-connected industrial site, transferable power contract, fiber route, or general-purpose laboratory may outlast the model that first filled it. These assets carry execution risk, and later efficiency gains can route around some of them. Their value is less dependent on one technical solution. That distinction explains how a real technology can produce overbuilding: firms invest early because waiting is strategically costly, and some projects become stranded even when the broader demand thesis proves right.

Place matters for the same reason. An API crosses a border in milliseconds; the cluster behind it cannot. Frontier systems remain attached to power, cooling, fiber, land, law, and ownership. Using co-location megawatts as its capacity proxy, the World Bank estimated that high-income countries held 77 percent of global data-center capacity as of June 2025, while low-income countries held less than 0.1 percent.[39] The benefits of cheap model access may spread much faster than the rents from owning the physical system.

Countries that import capable models but own few durable complements may receive better services without capturing much of the upside. The state shapes this distribution through infrastructure, procurement, standards, research funding, and the rules governing access to scarce inputs.

Before trying to govern recursive AI, states need a better view of what is happening inside the laboratories. Benchmark scores do not reveal whether accepted work is rising or whether development cycles are shortening. Large developers could report, confidentially and on comparable terms, AI-assisted changes that survive review, experiment throughput, reviewer time, research inference cost, and the duration of major development stages. Chan and coauthors propose a broader operational measurement framework along these lines.[2]

Confidentiality is not a novel obstacle. Nuclear safeguards use declarations, inspections, material accountancy, and independent validation while protecting sensitive information.[40] A UN scientific advisory report has outlined options for frontier AI that include third-party audits and monitoring of

software, hardware, or compute.[41] The aim would be a trusted aggregate picture, not publication of model weights or research plans.

Verification is productive infrastructure as well as oversight. A lab reviews code it cannot trust. A company keeps a human approver when liability is unclear. A regulator limits a system it cannot audit. Better evaluation, provenance, sandboxing, access control, and incident reporting can make it easier to justify admitting more machine output to ordinary use. NIST's generative-AI risk profile recommends independent testing, documented validity, continuing monitoring, and explicit uncertainty.[42]

Yet state capacity also includes deliberate delay. Courts need evidence and appeal; science needs replication; communities need a way to contest risks they will bear. The test is whether delay produces something: evidence exposed, responsibility assigned, an appeal made possible. An agency that merely cannot process an application is a bottleneck; review that earns legitimacy is part of the work. Machine-speed proposals will make the two easier to confuse.

WHAT WOULD CHANGE MY MIND

The number from the opening needs a causal interpretation. A six-month release interval does not show that AI cut six months from development. The lab may have used more compute, narrowed the target, changed its safety process, or timed a product launch. A credible estimate would compare teams or stages with different access to research agents, or decompose the cycle against a defensible baseline. It should count review, retries, integration, and inference rather than treating generated output as free.

By the end of 2027, I would want to see large laboratories report at least some comparable workflow evidence: accepted AI-assisted changes, experiment throughput, review time, and research inference cost. That date is my checkpoint, not a prediction borrowed from the cited literature. Disclosure may need to be confidential, but an independent party should be able to test the aggregate claim.

Stronger technical evidence would come from a system that chooses a multi-step research program on held-out problems and beats a strong human-designed search budget under an external evaluator. Open-research evaluations and RE-Bench point in this direction, but they do not yet establish frontier successor construction.[11],[13] The most persuasive result would be a retained algorithmic or organizational improvement that demonstrably pulls forward a broadly better successor, followed by that successor improving the same path again.

Another harness that edits itself on a fixed benchmark would be interesting without settling the issue.[4] The change must survive outside the environment that rewarded it. If local gains repeatedly vanish at frontier scale, transfer remains the brake.

The economic claims have separate tests. If AI becomes a material substitute for cognitive labor, highly exposed sectors should eventually show output or productivity separating from hiring, payroll, or labor income. If research inference becomes a major input, disclosed inference allocation should reveal it alongside spending and power demand. If capable models diffuse fast enough to erase frontier rents, concentration should decline even as use rises.

I would reduce my estimate of recursive feedback if task horizons keep lengthening while accepted R&D output stays flat. I would reduce it if machine-selected research continues to fail outside prepared environments, if review time rises as fast as candidate production, or if successor-development time fails to fall against a defensible baseline while benchmark scores continue to improve. Several of those results together would leave us with powerful automation rather than a self-sustaining research loop.

The stronger thesis begins when AI-assisted work is accepted, independently evaluated, and transferred across the organizational boundary in a way that shortens the successor cycle. A private lab may reach that point before outsiders can verify it, which makes measurement more urgent, not speculation more useful.

CONCLUSION

Until a laboratory can show against a credible baseline that one model helped build the next faster, recursive self-improvement remains a direction of travel. The public record shows pieces of the loop, not the loop closed.

REFERENCES

Suggested citation

The Recursive Economy: What Happens When Intelligence Becomes an Input into the Production of Intelligence? Version 1.0, published 15 August 2026. https://recursive-economy.pages.dev/

Publication note

This essay draws only on public sources and original scenario analysis; it claims no privileged access. Illustrative models are not forecasts or investment advice.

NOTES

1. Erik Brynjolfsson, Lorin M. Hitt, and Shinkyu Yang, "Intangible Assets: Computers and Organizational Capital," *Brookings Papers on Economic Activity* 2002, no. 1: 137–198. https://www.brookings.edu/articles/intangible-assets-computers-and-organizational-capital/ ↩

2. Alan Chan, Ranay Padarath, Joe Kwon, Hilary Greaves, and Markus Anderljung, "Measuring AI R&D Automation," arXiv:2603.03992 (2026). https://arxiv.org/abs/2603.03992 ↩1 ↩2

3. Google DeepMind, "AlphaEvolve: A Gemini-Powered Coding Agent for Designing Advanced Algorithms," 14 May 2025. https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/ ↩

4. Jenny Zhang et al., "Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents," arXiv:2505.22954 (2025), and Sakana AI project page. https://arxiv.org/abs/2505.22954 and https://sakana.ai/dgm/ ↩1 ↩2

5. Mingguang Chen, Licheng Wang, and Bo Qu, "Recursive Self-Improvement in AI: From Bounded Self-Refinement to Autonomous Research Loops," arXiv:2607.07663 (2026). https://arxiv.org/abs/2607.07663 ↩1 ↩2

6. METR, "Time Horizons," updated 2026. https://metr.org/time-horizons/ ↩

7. METR, "Measuring AI Ability to Complete Long Software Tasks," 19 March 2025. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/ ↩

8. METR, "Limitations of Time-Horizon Measurements," 22 January 2026. https://metr.org/notes/2026-01-22-time-horizon-limitations/ ↩

9. Joel Becker et al., "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity," METR, July 2025, with 24 February 2026 uplift update. https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/ and https://metr.org/blog/2026-02-24-uplift-update/ ↩1 ↩2

10. Chris Lu et al., "Towards End-to-End Automation of AI Research," *Nature* (2026), and Sakana AI, "The AI Scientist-v2," 2025. https://doi.org/10.1038/s41586-026-10265-5 and https://pub.sakana.ai/ai-scientist-v2/paper/ ↩

11. Peter Kirgis et al., "Can AI Agents Conduct Open-Ended AI Research? Early Evidence from Two Case Studies," arXiv:2607.27191 (2026). https://arxiv.org/abs/2607.27191 ↩1 ↩2 ↩3

12. Anthropic, "Automated Alignment Researchers: Using Large Language Models to Scale Scalable Oversight," 14 April 2026. Primary laboratory study. https://www.anthropic.com/research/automated-alignment-researchers ↩

13. Hjalmar Wijk et al., "RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts," arXiv:2411.15114 (2024). https://arxiv.org/abs/2411.15114 ↩1 ↩2

14. OpenAI, "PaperBench: Evaluating AI's Ability to Replicate AI Research," 2 April 2025. Primary laboratory study. https://openai.com/index/paperbench/ ↩

15. Gene M. Amdahl, "Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities," *AFIPS Conference Proceedings* 30 (1967): 483–485. https://doi.org/10.1145/1465482.1465560 ↩

16. Tom Davidson, Basil Halperin, Thomas Houlden, and Anton Korinek, "When Does Automating AI Research Produce Explosive Growth? Feedback Loops in Innovation Networks," NBER Working Paper 35155, April 2026. https://www.nber.org/papers/w35155 ↩

17. Tom Cunningham, Lukas Althoff, Basil Halperin, Brian Jabarian, Andrew Koh, Arjun Ramani, Phil Trammell, Parker Whitfill, and Cheryl Wu, "The Economics of Recursive Self-Improvement," Elasticity Institute working paper, 13 July 2026. https://elasticity.institute/rsi-paper.pdf ↩

18. Epoch AI, "LLM Inference Prices Have Fallen Rapidly but Unequally across Tasks," 12 March 2025. https://epoch.ai/data-insights/llm-inference-price-trends ↩

19. International Energy Agency, *Energy and AI*, 2025, executive summary and data-centre demand analysis. https://www.iea.org/reports/energy-and-ai/executive-summary ↩

20. Sarah Smith et al., *United States Data Center Energy Usage Report: 2025 Update*, Lawrence Berkeley National Laboratory, June 2026. https://eta-publications.lbl.gov/publications/united-states-data-center-energy-2025 ↩

21. Lawrence Berkeley National Laboratory, "Backlog of Power Plants Seeking Transmission Grid Connection Eased Somewhat in 2025 Amidst High Withdrawals," 1 July 2026. https://emp.lbl.gov/news/backlog-power-plants-seeking-transmission-grid-connection-eased-somewhat-2025-amidst ↩

22. Neel Somani, *Power 2026: Electricity Pricing in the Age of AI*, 2026. https://power2026.ai/ ↩

23. Stanford Institute for Human-Centered Artificial Intelligence, *Artificial Intelligence Index Report 2026*, April 2026. https://hai.stanford.edu/ai-index/2026-ai-index-report ↩

24. Anthropic, "Anthropic Economic Index Report: Economic Primitives," 15 January 2026. https://www.anthropic.com/research/anthropic-economic-index-january-2026-report ↩

25. US Federal Trade Commission, *AI Partnerships & Investments Study*, staff report, January 2025. https://www.ftc.gov/news-events/news/press-releases/2025/01/ftc-issues-staff-report-ai-partnerships-investments-study ↩

26. OECD, *Competition in Artificial Intelligence Infrastructure*, OECD Roundtables on Competition Policy Paper 330, 14 November 2025. https://doi.org/10.1787/623d1874-en ↩

27. Pawel Gmyrek et al., *Generative AI and Jobs: A Refined Global Index of Occupational Exposure*, International Labour Organization, 20 May 2025. https://doi.org/10.54394/HETP0387 ↩

28. US Census Bureau, "The Microstructure of AI Diffusion: Evidence from Firms, Business Functions, and Worker Tasks," Center for Economic Studies Working Paper CES-WP-26-25, April 2026. https://www.census.gov/library/working-papers/2026/adrm/CES-WP-26-25.html ↩

29. Anthropic, "Labor Market Impacts of AI: A New Measure and Early Evidence," 2026. https://www.anthropic.com/research/labor-market-impacts ↩

30. Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond, "Generative AI at Work," *Quarterly Journal of Economics* (2025). https://doi.org/10.1093/qje/qjae044 ↩

31. Shakked Noy and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science* 381 (2023): 187–192. https://doi.org/10.1126/science.adh2586 ↩

32. Anders Humlum and Emilie Vestergaard, "Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI," NBER Working Paper 33777, May 2025, revised March 2026. https://www.nber.org/papers/w33777 ↩

33. Fabrizio Dell'Acqua et al., "The Cybernetic Teammate: A Field Experiment on Generative AI and Teamwork," *Organization Science* (2026). https://doi.org/10.1287/orsc.2025.20702 ↩

34. Daron Acemoglu and Pascual Restrepo, "Automation and New Tasks: How Technology Displaces and Reinstates Labor," *Journal of Economic Perspectives* 33, no. 2 (2019): 3–30. https://doi.org/10.1257/jep.33.2.3 ↩

35. David Autor, Caroline Chin, Anna Salomons, and Bryan Seegmiller, "New Frontiers: The Origins and Content of New Work, 1940–2018," *Quarterly Journal of Economics* 139, no. 3 (2024): 1399–1465. https://doi.org/10.1093/qje/qjae008 ↩

36. Philip Trammell and Anton Korinek, "Economic Growth under Transformative AI," NBER Working Paper 31815, 2023, revised April 2026. https://www.nber.org/papers/w31815 ↩

37. International Labour Organization, *World Employment and Social Outlook: May 2025 Update*, 28 May 2025. https://www.ilo.org/publications/flagship-reports/world-employment-and-social-outlook-may-2025-update ↩

38. Alphabet, "2025 Q4 Earnings Call," 4 February 2026, and "2026 Q2 Earnings Call," 22 July 2026. Primary company disclosures. https://abc.xyz/investor/events/event-details/2026/2025-Q4-Earnings-Call-2026-Dr_C033hS6/default.aspx and https://abc.xyz/investor/events/event-details/2026/2026-Q2-Earnings-Call-2026-GgTAq7Is0z/default.aspx ↩

39. World Bank, *Digital Progress and Trends Report 2025: Strengthening AI Foundations*, disclosed 9 January 2026. https://doi.org/10.1596/978-1-4648-2264-3 ↩

40. International Atomic Energy Agency, *IAEA Safeguards Glossary: 2022 Edition*, 2022. https://www-pub.iaea.org/MTCD/Publications/PDF/PUB2003_web.pdf ↩

41. UN Secretary-General's Scientific Advisory Board, "Verification of Frontier AI Models," 13 June 2025. https://www.un.org/scientific-advisory-board/en/verification-frontier-ai-models ↩

42. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1, 26 July 2024; publication webpage updated 8 April 2026. https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence ↩