

Ataraxia

Nine Essays on Artificial Intelligence and the Economy

Shay O'Kelly

August 2026

R A X I A

A Note on the Name

Ataraxia is an old Greek word for a particular kind of calm. The earned equanimity of someone who has looked hard at how things actually work and stopped being rattled by them. The Epicureans held that you reach it through understanding. Fear comes from confusion about how the world works, and the cure is to study the machine until the noise loses its grip.

I chose it as the title because the decade these essays describe will be the loudest in economic history. Every year of the buildout will bring a bubble call, a doom headline, a crash scare, and an arms-race panic, and most of them will sound convincing. The only way to hold conviction through that, in a portfolio or just in a worldview, is to understand the underlying machine well enough that volatility reads as weather instead of judgment. These nine essays are my attempt to earn that calm by studying the capex, the bottlenecks, the geopolitics, and the distribution fights until the conclusion stops being exciting and starts being obvious. Calm is not the absence of conviction. It is what conviction feels like once the work is done.

It is also where my fund, RAXIA, takes its name.

Contents

I. Manufactured Intelligence: Why AI Is Not a Technology Story	4
II. The Factories: What It Costs to Build Intelligence.....	7
III. The Macroeconomics of AI: Cost Savings vs. Explosive Output Growth.....	12
IV. The Bottlenecks: Where Value Accrues	16
V. The Arms Race: Geopolitics of the Buildout	21
VI. The Ledger: Wealth and Work in the Machine Economy.....	26
VII. The Wildcard: Alignment	30
VIII. The Future of Tokens: Demand at Machine Speed	34
IX. The Opportunity: Crowded Theme, Empty Trade	38

Manufactured Intelligence: Why AI Is Not a Technology Story

For all of human history, intelligence has been the one economic input you could not buy more of. Energy scaled, first with wood, then coal, then oil, then the grid. Land scaled with conquest and agriculture. Labor scaled with population. Capital scaled with finance. But cognition, the thing that actually invents everything else, only ever scaled at the speed of biology. If you wanted more thinking in the world, you needed more people, raised over decades and educated over decades more. Every civilization, company, and research lab has operated under that constraint, so completely that we stopped noticing it was a constraint at all. AI turns intelligence into a manufactured good, with a price, a supply curve, and a capex line. Framed that way, the story is bigger than technology. It removes what has always been the binding constraint on growth, which is why I would put it with industrialization, maybe even agriculture, on the short list of true economic regime changes.

To see why, it helps to notice that this has happened twice before, in slower motion. Intelligence has gone through three regimes. The first was biological. For millions of years, cognitive capacity improved at the pace of evolution, and knowledge died with the brain that held it. The second regime was cultural. Language, then writing, then printing let knowledge outlive its host and accumulate across generations. This is why economic growth for most of history tracked population. More brains meant more ideas, and better information technology meant less rediscovery of what someone had already figured out. The printing press did not make any individual smarter, but it changed the replication economics of knowledge, and within a few centuries came the scientific revolution and then the industrial one. Each regime change was not an intelligence upgrade. It was a bandwidth upgrade. Which is what makes the third regime so violent as a discontinuity. A trained model is cognition itself, not a description of it, and it copies at essentially zero marginal cost. Culture let us copy what a mind had learned. Synthetic intelligence lets us copy the mind.

The standard history of AI, from the 1956 Dartmouth workshop through the winters of the 70s and 80s to the deep learning breakthrough of 2012¹, is usually told as a story of scientific dead ends and eventual insight. I see it differently. The core ideas won almost nothing on novelty. Neural networks date to the 1940s and 50s, backpropagation was published in usable form in 1986², and the researchers who carried connectionism through the winters were mostly right the whole time. What they were missing was nine orders of magnitude of compute cost decline. The AI winters look less like failures of imagination than long stretches where the hardware economics had not caught up to hypotheses sitting in the literature for decades. When compute got cheap enough, in GPU clusters originally built to render video games, the old ideas started working almost immediately. That reading matters because it tells you what kind of problem intelligence turned out to be. It was never a mystery waiting for a genius. It was an industrial input waiting for its cost curve.

The scaling laws made that explicit. Starting around 2020, researchers showed that model capability improves as a smooth, predictable function of compute, data, and parameters, holding across many orders of magnitude³. It is hard to overstate how strange and how important that is. Capability in every prior technology came from invention, which is lumpy and unschedulable. Capability in AI comes, to a first approximation, from spend. The scaling laws turned artificial intelligence from a research problem into a procurement problem, and the industry noticed. The defining questions in AI today are megawatts, fab allocation, and financing structures, which is why the org charts of AI labs increasingly resemble energy companies. When the smartest thing on earth improves predictably with dollars, the interesting questions all become questions about dollars, land, power, and silicon.

So where is this heading? My view, and this is the claim the rest of these essays build on, is that both capability and diffusion will outrun what consensus assumes, for two separate reasons. On capability, the scaling laws have not bent, and the newer ingredient, letting models think longer and act as agents rather than answer single prompts, compounds on top of pretraining rather than replacing it. Nothing in the data says we are near a ceiling. On diffusion, I think Wall Street is using the wrong reference class. Consensus models AI adoption like enterprise software, a slow S-curve gated by IT budgets and change management. But when what you are buying is labor rather than a tool, adoption is not gated by budget cycles, it is gated by the cost delta, and the cost delta between a knowledge worker and an API call is not twenty percent, it is often a hundred to one. Even if capability froze today, current models are so under-deployed relative to what they can already do that diffusion alone locks in years of growth. The gap between what already exists and what is actually used may be the largest unpriced fact in the market. And capability is not freezing.

None of this requires believing in imminent superintelligence. It requires believing three things that I think are now well-evidenced. Intelligence has become manufacturable, its quality improves predictably with investment, and its output is a substitute for the most expensive input in the economy. If those hold, then the constraint on growth shifts from how many smart people exist to how much intelligence-producing capital we can build. Which turns the most important question in economics into an embarrassingly physical one, what does it actually cost to build the factories? That is the subject of the next essay, and the numbers are bigger than almost anyone is ready for.

Sources

1. The Dartmouth Summer Research Project on Artificial Intelligence (1956), and Krizhevsky, Sutskever, and Hinton, ImageNet classification with deep convolutional neural networks (2012).
2. Rumelhart, Hinton, and Williams, "Learning representations by back-propagating errors," Nature, 1986.
3. Kaplan et al., "Scaling Laws for Neural Language Models" (2020), and Hoffmann et al., "Training Compute-Optimal Large Language Models" (2022).

The Factories: What It Costs to Build Intelligence

If the last essay was right that intelligence has become a manufactured good, then the obvious next question is what the factories cost, and the answer is that we are living through the largest infrastructure buildout in the history of capitalism. The four big hyperscalers alone, Amazon, Google, Meta, and Microsoft, have guided to roughly \$725 billion of capital expenditure for 2026, up 77% from about \$410 billion the year before¹. Add OpenAI's \$500 billion Stargate program, the sovereign projects, and the neocloud buildouts, and a trillion dollars a year is within sight. In the first quarter of 2026, AI data centers, hardware, and networking amounted to 1.4% of US GDP and drove roughly three quarters of the economy's growth². Relative to GDP this is already bigger than the telecom buildout at the peak of the dot-com era, and on the bullish forecasts it reaches 4% of GDP, which is railroad-boom territory³. That is the right comparison, and not as hyperbole. This is the canal-and-railroad phase of an industrial revolution. Enormous fixed investment, deployed ahead of the revenue, carrying the whole economy while it happens.

But the railroad analogy breaks in one place, and the break is where all the interesting questions live. The money is buying two completely different kinds of assets that get lumped into one capex line. The first is the durable layer: land, buildings, cooling, substations, transmission, and power generation, assets that will run for thirty to fifty years no matter whose chips sit inside. The second is the consumable layer, the chips themselves, which are roughly half the spend and carry accounting lives of three to six years. A railroad was bought once and used for a century. GPUs have to be repurchased forever. Which means intelligence does not have a construction cost, it has an ongoing production cost, like electricity. The capex line is really an opex line in disguise, and that is exactly what you would expect once intelligence becomes a manufactured good rather than a built one.

The consumable layer is where the bears attack, and the attack deserves to be taken seriously. Michael Burry and others have accused the hyperscalers of overstating earnings by stretching GPU depreciation schedules⁴, and the accounting point is real. Small changes in assumed useful life shift billions of dollars of reported profit without changing a dollar of cash. If a chip is economically dead in three years but depreciated over six, the industry's earnings are inflated and the buildout is more fragile than it looks.

Here is why I think the bears have the depreciation story backwards. Look at what old chips actually earn. The H100 launched in 2022 and rented for as much as \$8 an hour at the peak of the 2023 shortage. By late 2025 the price had collapsed to around \$1.70, which is the moment the depreciation bears point to. But then something happened that their model says should not happen, the price went back up, rising almost 40% to about \$2.35 by March 2026, because on-demand H100 capacity is effectively sold out across the market and renters who locked in capacity refuse to release it⁵. A four-year-old chip, two full generations behind the frontier, is getting more expensive to rent. The reason is that training and inference are different businesses. Training the next frontier model demands the newest silicon, but serving models to users does

not, and modern mixture-of-experts architectures only activate a fraction of their parameters per query, which makes older chips perfectly viable inference machines. So the installed base does not retire when a new generation ships. It migrates from training to inference and stays full. And the economics of that inference work are absurd by any labor standard. When a machine does fifty dollars an hour of work for two dollars, someone will always rent it.

There is also a supply-side reason obsolescence runs slower than the accounting debate assumes. Every leading-edge chip on earth passes through the EUV lithography machines of a single company, ASML, which will ship about 65 of them in 2026, up from 44 last year, and perhaps 85 next year⁶. Not thousands. Tens. The entire industry's ability to replace old chips with new ones is rationed by a production line measured in dozens of machines a year, which is why Blackwell lead times stretched into mid-2026 even as Nvidia ramped as fast as it could. Depreciation is not really a function of a chip's age. It is a function of replacement supply, and when replacement is rationed, old capital stays economically alive far longer than the spreadsheet says. In a supply-constrained world, the choice is not between old chips and new chips. It is between old chips and no chips.

The other thing changing is who writes the checks. The first phase of the buildout was funded from hyperscaler operating cash flow, arguably the safest financing in corporate history. That era is ending. Meta's Hyperion campus in Louisiana was financed through a \$30 billion special purpose vehicle, with \$27 billion of A+ rated debt anchored by PIMCO and BlackRock, the largest private credit deal ever done⁷. Stargate runs on stacked SPV debt from Blue Owl, JPMorgan, and others. Estimates of off-balance-sheet AI debt now run north of a trillion dollars, and senators on the Banking Committee have started writing letters about it⁸. I would resist the urge to treat this as a scandal, because it is how every great buildout in history was financed. The railroads ran on bonds too. But it does move risk from equity holders who can absorb losses to credit markets that have historically been bad at underwriting technology obsolescence. That leaves long-dated paper against assets whose useful life is the single most contested number in the market, and the tension is real. The old-chip economics above suggest the collateral lives longer than the bears claim. It does not suggest the debt is riskless.

The sharpest version of the financing critique is circularity. Nvidia invests in OpenAI and CoreWeave, which use the money to buy Nvidia chips, and Nvidia books the revenue. AMD handed OpenAI warrants on 160 million shares priced at a penny each, vesting as OpenAI buys AMD GPUs. By this summer, Nvidia was reportedly discussing guaranteeing up to \$250 billion of OpenAI's data center lease payments and financing another \$350 billion of chip purchases⁹. Bears call this vendor financing, and they are right. That is exactly what it is. It is the same structure that let Lucent and Nortel manufacture a demand boom in 1999, and the concern has escalated from an analyst complaint to an official one. The Bank for International Settlements now lists circular AI financing among the biggest risks to global financial stability¹⁰.

It matters, though, to separate two things the bears lump together. An equity stake in a customer is self-limiting. If Nvidia puts two billion dollars into CoreWeave and CoreWeave fails, Nvidia loses two billion dollars, takes a write-down, and moves on. A guarantee is different in kind, not in size. If Nvidia guarantees \$250 billion of OpenAI's lease payments and OpenAI's revenue stalls, those obligations do not disappear with the customer. They land on Nvidia. The lenders who financed those data centers get paid by Nvidia or they do not get paid at all, and their debt is not held by venture capitalists who priced in failure. It is rated A+ and sits with insurers, pension funds, and credit funds that bought it as a safe investment. Equity losses stop with the holder. Credit losses cascade, because the holders are themselves leveraged, and a failure would not just slow the boom, it would ripple through the bond market from inside the most important company in the world. So my rule for reading these deals is simple. The investments are noise, the guarantees are the risk. Watch the guarantees.

But here is what the circularity debate keeps missing, and where I part ways with the bears entirely. Vendor financing is only fatal when the end customer never shows up. That is what actually killed the telecom vendors. They financed carriers building networks for traffic that did not exist, so when the lending stopped there was no revenue under any layer of the stack. Which means the whole question collapses to one measurable thing. Is there real, paying, arms-length demand at the end of the loop? That is no longer a matter of opinion. Google processed 3.2 quadrillion tokens in May, seven times more than a year earlier and more than three hundred times the level of two years ago¹¹. Microsoft was running over 100 trillion tokens a quarter, five times its prior year, as far back as the spring of 2025¹². Token consumption is already running at more than double what the equipment makers were forecasting for 2028. On revenue, Anthropic went from one billion to forty-seven billion dollars of annualized revenue in less than a year and a half, the fastest revenue scaling in corporate history, with OpenAI around twenty-five billion¹³, and that money comes overwhelmingly from businesses paying market prices, not from Nvidia's balance sheet. Nobody is subsidizing the two dollar an hour H100 renter, and the market for those four-year-old chips is sold out anyway. The sold-out market for old chips is the fifty-for-two trade already happening at scale. So my bull case is simple. The loop is real, and it is being lapped by demand. Compute consumption is compounding at several hundred percent a year while the physical capacity to serve it grows at fifty. The vendor financing exists not because demand is missing, but because demand is arriving faster than any balance sheet can build for it. The 1999 comparison fails on the only fact that matters. This time the end customer showed up. What we have is a shortage wearing a bubble costume.

Which leaves the ceiling question. How big can the demand actually get? The bear math stacks a trillion of annual capex against AI revenues measured in the low hundreds of billions and calls it a bubble. I think the reference class is wrong, the same way it was wrong in the last essay. This infrastructure is not underwritten against software budgets. It is underwritten against the global wage bill for cognitive work, which runs to tens of trillions of dollars a year. Do the division. A trillion dollars of annual infrastructure spend needs to capture only low single digits of that wage

bill to clear its cost of capital. Priced against SaaS, the buildout looks indefensible. Priced against labor, it looks conservative.

So the checks are being written, the debt is being raised, and the spending already shows up in national output as the largest single driver of American growth. And yet consensus forecasts for the decade ahead barely move. The same two percent growth, the same productivity trend, as if a railroad-scale buildout were a rounding error. Someone is badly wrong, and the next essay is about why I think it is the forecasters, not the builders.

Sources

1. Hyperscaler 2026 capex guidance (~\$725B combined, up 77% from ~\$410B). Company guidance compiled by ValueAdd VC, CreditSights, and Yahoo Finance.
2. AI capex at 1.4% of US GDP and ~75% of Q1 2026 growth. Via Paul Kedrosky, "Honey, AI Capex is Eating the Economy," Epoch AI, and TECHi.
3. Railroad and telecom era comparisons as a share of GDP. Via Paul Kedrosky and Epoch AI.
4. GPU depreciation debate and stretched useful lives. Deep Quarry via National Law Review, and Michael Burry public commentary.
5. H100 rental price history (\$8 peak 2023, ~\$1.70 October 2025, ~\$2.35 March 2026, sold-out on-demand capacity). Via GPUSmith and IntuitionLabs.
6. ASML EUV shipments (44 in 2025, ~65 in 2026, 80-85 guided 2027). ASML guidance via TechPowerUp and CryptoBriefing.
7. Meta Hyperion \$30B SPV with \$27B A+ debt (PIMCO, BlackRock). Via Global Datacenter Hub and Capacity.
8. Off-balance-sheet AI debt above \$1T, and the US Senate Banking Committee letter (January 2026). Via Techerati and the US Senate Banking Committee.
9. Nvidia-OpenAI discussions (\$250B lease guarantees, \$350B chip financing). Wall Street Journal and Bloomberg reporting, July 2026. Nvidia \$2B CoreWeave investment plus \$6.3B capacity agreement, and AMD-OpenAI 160M share warrants. Via Yahoo Finance and EBC Financial Group.
10. BIS 2026 Annual Report naming circular AI financing a top financial stability risk. Bank for International Settlements.
11. Google 3.2 quadrillion tokens per month (7x YoY). Google I/O 2026.
12. Microsoft 100T+ tokens per quarter (5x YoY). Microsoft FY2025 Q3 earnings call, April 30, 2025.
13. Anthropic ~\$47B and OpenAI ~\$25B annualized revenue. Via Trending Topics, ValueAdd VC, and Epoch AI.

The Macroeconomics of AI: Cost Savings vs. Explosive Output Growth

The most cited macroeconomic estimate of AI's impact is also probably the most misleading. In 2024, Daron Acemoglu ran the numbers on generative AI and concluded that it would add about 0.71% to total factor productivity over ten years, which works out to roughly 0.07% per year¹. Even after accounting for the extra capital investment AI would induce, he lands on a total GDP boost of somewhere between 0.9% and 1.8% over a decade. That is a rounding error. If he is right, AI is basically a slightly better version of Excel. I think he is wrong, and the reason he is wrong is not that his math is bad. It is that the framework itself quietly assumes away the most important possibility, that AI is not a tool that helps workers, but a new kind of worker altogether.

Start with how Acemoglu actually gets his number, because the mechanics matter. He uses Hulten's Theorem², which says that the aggregate TFP gain from a technology equals the share of economic tasks exposed to it multiplied by the average cost savings on those tasks. His inputs are reasonable on their face. About 19.9% of U.S. labor tasks are exposed to generative AI. Of those, only about 23% are cost-effective to automate within ten years, which leaves 4.6% of total tasks actually affected. On each affected task, he estimates 27% labor cost savings, and since labor is about 57% of total costs, that translates into roughly 15.4% total cost savings per task. Multiply 4.6% by 15.4% and you get the famous 0.71%.

Every step of that chain is defensible, and the conclusion is still deeply misleading, because the whole exercise only measures one thing, the intensive margin. It asks how much cheaper it gets to do the tasks the economy already does, inside the firms that already exist. It holds the task matrix fixed. It assumes nobody restructures a workflow, nobody launches a company that was previously impossible to staff, and nobody responds to a collapse in the price of cognitive work by simply consuming a lot more of it. That last omission is the Jevons Paradox, and it is not a fringe concern. When the price of a valuable input falls dramatically, demand for it tends to explode, and total spending often goes up rather than down. Cheaper drug discovery does not mean pharma spends less on discovery. It means someone finally screens the millions of molecules that were never financially viable to test. Hulten's Theorem is built for small perturbations around a steady state. It is not built for a world where an entire class of inputs drops 99% in price.

The deeper problem sits in the accounting. National accounts, as kept by the BEA, define labor strictly as human hours worked. AI and software are capital, full stop. So if a thousand AI agents execute a software project end to end, measured labor growth is exactly zero, and the output shows up as capital income and corporate profits rather than wages. This is not just a bookkeeping quirk. It smuggles a substantive economic assumption into the model. In the standard Cobb-Douglas production function, $Y = A \cdot K^{\alpha} \cdot L^{1-\alpha}$, with capital share α around 0.35, adding more capital runs into diminishing returns almost immediately, because the fixed supply

of human labor is the binding constraint. Give a worker a second laptop and you get very little. The entire structure of the model guarantees that “more K” cannot generate explosive growth.

But that logic breaks the moment capital can do cognitive work autonomously. Adding another GPU cluster is no longer handing a human a better tool. It is adding a cloned worker who runs 24/7, never quits, and can be replicated at the speed of a data center buildout. When capital and labor become near-perfect substitutes, the production function drifts from $Y = A \cdot K^\alpha \cdot L^{1-\alpha}$ toward something much closer to $Y = A \cdot K$, where capital now includes both physical hardware and replicable digital labor. In an AK regime, the diminishing returns that anchor the Solow model to 2% growth simply stop binding, and growth goes from linear to compounding.

You can see what this means by decomposing growth the standard way: $g_Y = g_A + \alpha \cdot g_K + (1 - \alpha) \cdot g_L$. In the baseline peacetime economy, TFP grows around 1%, capital around 2.5%, human labor around 0.5%³, and you get the familiar 2% to 2.5% GDP growth. In a five-year AI acceleration scenario, TFP picks up to maybe 1.5% to 2.5%, capital growth doubles as the compute buildout scales, the capital share drifts toward 0.45 or 0.50, and GDP growth lands in the 3.5% to 5.5% range. Noticeable, but still recognizable. The regime break comes further out. In a full AK transition, where the capital share climbs toward 0.80 because cognitive labor is fully software-driven, TFP growth of 15% to 20% combined with capital growth of 20% to 50% produces something like 30% annual GDP growth, which is an economy doubling every two and a half years or so. That sounds insane by historical standards, but nothing in the arithmetic forbids it. What forbade it before was the human labor bottleneck, and that is precisely the thing that autonomous digital labor removes.

Which raises the obvious question. If human labor stops being the constraint, what is? The answer is physical. Some will argue the truly scarce input stays human, taste and judgment about what is worth building, and I take that seriously, but taste does not gate throughput the way megawatts do, and this essay is about throughput. The bottleneck migrates from bits to atoms. Energy is the clearest one, since gigawatt-scale power generation and grid capacity become the effective speed limit on how much cognition the economy can run. Then hardware and materials: silicon fabs, specialized chips, lithium, copper, cooling. And finally physical experimentation itself, because software iterates in seconds while clinical trials, materials synthesis, and manufacturing buildouts run on feedback loops measured in months and years. If something like superintelligence shows up on a short horizon, say three years, the ceiling on software efficiency effectively disappears and macroeconomics stops being a study of human capital allocation. It becomes a study of thermodynamics and permitting.

There is one more consequence worth taking seriously, and it is distributional. As the wage share of income falls, two things happen at once. Knowledge work undergoes hyper-deflation, with the marginal cost of legal work, diagnostics, engineering design, and software heading toward zero, which is genuinely great for real purchasing power. But the fiscal system runs on payroll and income taxes, which means the tax base erodes exactly as capital income explodes. And markets

themselves face an underconsumption problem if households lose wage income faster than they gain access to cheap goods. The likely endpoint is a restructuring of how governments raise and recycle revenue. Compute taxes, heavier capital gains taxation, sovereign equity stakes in compute infrastructure, or direct dividends that push capital returns back into consumer demand. None of these are radical ideas once you accept the premise. They are just what fiscal policy looks like when the thing generating income is a data center instead of a payroll.

So the real debate is not about Acemoglu's parameter estimates. It is about a classification decision. Is AI capital or is it labor? The 2024-era models treated it as capital, a tool that makes humans marginally more productive, and a rounding error is all a tool can ever deliver, so the conclusion was baked into the assumption. The evidence since has been eating that assumption, agents doing unsupervised work, adoption behaving like labor substitution rather than software sales, demand compounding on machine timescales. If AI is labor, replicable, autonomous, and elastic in supply, then models built on a fixed task matrix and a human bottleneck are measuring the wrong thing entirely, and growth ends up bounded not by how many people we have, but by how much energy and silicon we can bring online. My bet is on the labor story, because that is where the constraint actually lives, and the rest of this series is the evidence.

Sources

1. Daron Acemoglu, "The Simple Macroeconomics of AI," 2024 (MIT, NBER Working Paper 32487). All baseline figures in this essay (19.9% task exposure, 23% feasible adoption, 27% cost savings, 0.71% ten-year TFP gain) are from this paper.
2. Charles Hulten, "Growth Accounting with Intermediate Inputs," Review of Economic Studies, 1978.
3. Standard growth decomposition values from BEA national accounts and BLS labor share data.

The Bottlenecks: Where Value Accrues

The first three essays made an argument for scale. Intelligence has become a manufactured good, the factories cost a trillion dollars a year, and the economic models underestimating what comes next are broken in identifiable ways. But scale is not an investment thesis. Growth and returns are different things, and the history of technology is a graveyard of investors who confused them. Airlines transformed civilization and destroyed capital for a century. The question that matters for money is never whether something is important. It is where the scarcity sits, because value does not accrue to importance. It accrues to whatever stays scarce while everything around it grows.

The lesson everyone half-remembers here is fiber. In the late 1990s, telecom companies raised hundreds of billions and buried tens of millions of miles of fiber optic cable on the thesis that internet traffic would explode. The thesis was right. The traffic came, and the investors were wiped out anyway, because supply outran demand so badly that bandwidth prices collapsed more than 90%¹ and most of the fiber sat dark for a decade. The value did not disappear. It fled up the stack, to the companies that consumed the glutted resource, Google, Netflix, and everyone else who built businesses on top of bandwidth that had become too cheap to meter. This is the template in every consensus forecast today. Infrastructure commoditizes, applications win, so fade the buildout and buy the software. Whether that template applies is, I think, the single most important investment question of the decade. And I think it does not, at least not on the timeline consensus assumes, for a reason you can verify layer by layer.

There is a structural reason to think the template inverts this time, and it comes back to the first essay. The internet economy was defined by operating leverage. Software cost essentially nothing to replicate, bandwidth glutted, and so value pooled in the one layer where marginal cost was zero, the application. Infrastructure was a commodity input to somebody else's margin. Intelligence breaks that logic, because intelligence is the first digital product with a real marginal cost. Every token is paid for in silicon, memory, and electricity. The thing being sold is not code that copies for free, it is the metered output of physical capital, which is exactly what the first essay meant by intelligence becoming a manufactured good. When the product itself consumes scarce physical capacity with every unit sold, operating leverage migrates down the stack, away from the software that spends tokens and toward whoever owns the capacity that makes them. The internet moved value up the stack because infrastructure was abundant. AI moves it down, because for the first time since the industrial era, the binding constraint on the digital economy is physical.

A glut requires supply to outrun demand. The demand side we covered in the last essay. Token consumption is compounding at several hundred percent a year, already double what equipment makers forecast for 2028². And even that understates it, because what we are watching is two exponential curves stacked on top of each other, running on different clocks. Model capability is already well up the steep part of its curve, compounding through scale and now through models

that think longer and act as agents. Diffusion, the share of the economy actually using this stuff, is still at the flat beginning of its own curve. Most companies are running pilots, most workflows are untouched, and agents barely exist in production. Adoption curves always look flat right before they go vertical. When diffusion hits its ramp while capability is still compounding, the two multiply, and today's token numbers, the quadrillions that already broke every forecast, will look like the quiet part. Inference is on its way to becoming the largest market in the world. So the question reduces to supply, and the receipts say supply is rationed at every layer of the stack. Walk it from silicon to power.

Start with silicon. Every leading-edge chip flows through ASML's EUV machines, roughly 65 shipping this year³. TSMC's advanced packaging capacity, the CoWoS lines that bolt GPU dies to their memory, is sold out through 2026 with lead times stretching deep into 2027, despite capacity roughly quadrupling in two years, from the mid 30,000s of wafers a month to a planned 130,000 by the end of this year, and still falling short of demand⁴. High-bandwidth memory is the same story. Samsung, SK Hynix, and Micron have sold out their entire output for the year ahead, and SK Hynix concedes 2026 is almost fully allocated⁵. Sit with what that means. The capacity expansions are selling out before they are built. In fiber terms, it is as if every mile of cable had been leased before it was buried. That is not how gluts form. That is how shortages persist.

Within silicon, memory deserves its own paragraph, because the market is mispricing a structural change as another cycle. Model architecture is doing something specific to hardware. Mixture-of-experts designs and relentless algorithmic efficiency keep cutting the compute needed per token, while total parameter counts keep climbing as models get more precise and have more training compute behind them, and context windows keep stretching as agents hold more in working memory. Less compute per token, with more parameters and context to hold, is a formula for inference becoming memory-bound. The AI server is quietly turning into a memory product with a logic chip attached, and HBM4, ramping now for the next accelerator generation, pushes memory's share of the bill of materials higher again⁶. Yet the memory makers are still priced like the pure commodity cyclical they were for thirty years, valued as if today's demand is a peak that mean-reverts on schedule. Given what the architecture curve is doing to memory intensity, I think those valuations are laughable. The market is pricing the most structurally advantaged layer of the stack off its most traumatic historical pattern.

Now power, which everyone, including the government, calls the great bottleneck. The numbers are real. GE Vernova's turbine backlog hit 116 gigawatts, production slots are sold out through 2029⁷, grid transformers that took a year before the pandemic now average nearly three, with worst cases toward five⁸, and of the twelve gigawatts of US data center capacity announced for 2026, only five are actually under construction⁹. But here I hold a minority view. Power is a softer bottleneck than the market believes, because it has workarounds and the hard bottlenecks do not. Power is a modest share of a data center project's cost, perhaps a tenth. And there are many ways around it. You can site campuses where stranded power already exists, build behind-

the-meter generation and skip the interconnection queue entirely, restart nuclear plants, deploy solar and storage in the desert, or shift flexible training workloads to off-peak hours. Every one of those paths is being used right now, and the turbine backlog is partly just the receipt for the workarounds being bought. Power is a real constraint that raises costs and adds years. It is not a wall. There is no equivalent list for lithography. Nobody sites around ASML. Nobody builds behind-the-meter TSMC. The semiconductor chain is a bottleneck in the strict sense, single points of failure, sold out for years, with expansion gated by physics and decades of accumulated process knowledge that no amount of capital can shortcut. When I rank where value accrues, I rank by substitutability, and the scarcest thing in the world right now is not a megawatt. It is a CoWoS slot.

This points to the principle that organizes the whole investment landscape. The bottleneck migrates, and value migrates with it. In 2023 and 2024 the constraint was GPUs themselves, and essentially all the economics of the boom pooled at Nvidia. By this year it had spread to memory and packaging, and businesses Wall Street priced as commodities for thirty years started printing record numbers with multi-year visibility. Power equipment is having its moment now, with backlogs stretching past 2029. But migration is only half the story, because each layer keeps its pricing power only as long as it stays scarce, and layers differ in how long they can stay scarce. Constraints with substitutes get engineered around in a few years. Constraints without substitutes hold for as long as demand holds. That is why I weight the semiconductor chain over energy even as energy gets the headlines. The returns are made by owning the constraint that cannot be routed around, before consensus stops pricing it as a cycle.

It is just as important to be clear about what is not scarce, and the uncomfortable answer is intelligence itself. The price of a token of a given capability has been collapsing continuously, open-weight models trail the frontier by months, and the labs compete each other's margins away even at extraordinary revenue scale. The biggest of them run rates in the tens of billions and still lose money on every dollar. The product is miraculous and the economics of producing it are brutal, which is the oldest combination in technology. On a decade horizon I would push this further. I believe everything at the application layer eventually commoditizes and gets priced toward the cost of tokens, because any workflow built on rented intelligence can be rebuilt by anyone else renting the same intelligence. And I believe the models themselves end the same way. Once models are doing the research and building the next generation of models, the advantage of any one lab's researchers stops compounding, frontier capability converges, and open-source implementations run everything toward the marginal cost of inference. The endgame of manufactured intelligence is intelligence priced like a utility. The one scenario that stops it is the big one. If AI becomes a first-order geopolitical concern, governments wall off frontier capability the way they wall off enriched uranium, and commoditization stops at national borders. I think that concern is real, and the next essay argues it is closer to the base case than the exception. Either way, the conclusion for where value accrues is the same. The durable rents are not in the layer where competition is a price war between geniuses, and not in applications being

priced toward token cost. They are in the layers underneath, where supply is rationed by machine counts, backlogs, and permits, and where no amount of brilliance lets a competitor conjure a turbine slot before 2030.

Which brings me to the mispricing, because a bottleneck everyone can see should already be in the price, and the reason I think it is not comes down to how Wall Street models these businesses. Turbine makers, memory companies, electrical equipment suppliers, and utilities are all priced as cyclical. The models assume today's demand is a peak, margins mean-revert, and backlogs are a moment in a cycle rather than a change in regime. That assumption is exactly what the first essay argued against. If model capability and diffusion keep outrunning consensus, then compute demand estimates are meaningfully too low, and what looks like the top of a cycle is actually the early innings of a permanent step-change in demand for power and silicon. The gap between those two readings, priced as a cycle, behaving like a regime, is where the returns live. It is the gap I am positioning for, long the bottlenecks, without leverage, because the thesis needs time more than it needs timing.

One more thing follows from all of this, and it opens the next essay. When a handful of chokepoints, one lithography company, one leading foundry, a few turbine makers, and the electrical grid, decide which nations get to manufacture intelligence and how fast, those chokepoints stop being industrial assets and become strategic ones. Governments have noticed. The buildout is already reorganizing alliances, export rules, and industrial policy, and that collision between the bottlenecks and geopolitics is where we go next.

Sources

1. Dot-com fiber overbuild and >90% bandwidth price collapse. Widely documented telecom history (McKinsey and FCC retrospectives on the 1996-2002 buildout).
2. Token consumption ~2.4x Dell's 2028 forecast, and Goldman Sachs projecting ~70x token growth 2025-2030. Via IO Fund and Goldman Sachs Research.
3. ASML EUV shipments (~65 in 2026). ASML guidance via TechPowerUp.
4. TSMC CoWoS sold out through 2026 with lead times of 52-78 weeks into 2027, and capacity from ~35k wafers/month (end-2024) to ~125-130k planned by end of 2026. TSMC shareholder meeting statements via TrendForce and DigiTimes, and SemiAnalysis.
5. HBM output sold out at Samsung, SK Hynix, and Micron, with SK Hynix 2026 allocation. Via AI CERTs, TrendForce, and company earnings commentary.
6. HBM4 ramp and rising memory share of accelerator bill of materials. Via SemiAnalysis and Wing VC, "The Memory Triopoly."
7. GE Vernova 116 GW turbine backlog, slots sold out through 2029, ~10 GW remaining across 2029-30, 5-7 year waits. Via Turbomachinery Magazine and Power Engineering.
8. Grid transformer lead times averaging ~2.5-3 years vs ~12 months pre-pandemic, worst cases of 4-5 years. Wood Mackenzie survey via T&D World, and The Conversation.
9. 5 GW under construction of 12 GW announced 2026 US data center capacity. Sightline Climate via Tech Fund.

The Arms Race: Geopolitics of the Buildout

The last essay ended on an observation that deserves its own essay. When a handful of physical chokepoints decide which nations get to manufacture intelligence and how fast, those chokepoints stop being industrial assets and become strategic ones. The twentieth century organized its geopolitics around oil, and we fought wars, built alliances, and drew maps around the places it came out of the ground. Intelligence is now on the same path, except the concentration is more extreme than oil ever was. There was never a moment when one company in one country made every barrel, but there is exactly one company on earth that makes the lithography machines behind every advanced chip, one foundry that manufactures nearly all of them, and it sits on an island claimed by the country America is racing. When national growth rates depend on a supply chain this concentrated, foreign policy becomes supply chain management, and that is not a metaphor. It is a literal description of what export controls are.

On paper, the chokepoint map is a Western royal flush. ASML is Dutch. TSMC is Taiwanese and fabs mostly for American designers. The frontier labs are American, the hyperscaler capital is American, and the high-bandwidth memory is Korean. Starting in 2022, Washington began playing that hand through export controls designed to deny China advanced chips and the tools to make them, and the logic followed directly from the scaling laws in the first essay. If capability is a function of compute, and you control compute, you control capability. It is the most aggressive use of economic chokepoints since the Cold War, and it rests on one assumption, that the denied input cannot be substituted or engineered around.

The evidence since then says the assumption is only half true, and the half that failed is instructive. In January 2025, DeepSeek released a reasoning model trained under embargo, on restricted hardware, that matched the leading Western models at a fraction of the training cost, and Nvidia lost almost six hundred billion dollars of market value in a single day processing the news, the largest one-day loss in market history¹. Since then the pattern has repeated on a schedule. Moonshot's Kimi K3, released this July at 2.8 trillion parameters, is the largest open-source model in the world². Eight of the top ten Chinese models are open-weight, and Alibaba's Qwen family has passed Meta's Llama in cumulative downloads to become the default open model for much of the world's developers³. The embargo was supposed to ration capability, but capability is a function of compute and ideas, and rationing one raises the return on the other. Constraint became a forcing function. Chinese labs got world-class at efficiency precisely because they were not allowed to be world-class at scale.

And notice what China does with those models, it gives them away. This is not idealism. It is the logic of the last essay weaponized. If you cannot monopolize the frontier, commoditize it, because every workflow in the world that runs on a free Chinese open-weight model is a workflow that generates no revenue for the American labs trying to earn back a trillion dollars of capex, and a workflow that runs on weights whose training data and assumptions were shaped in

Beijing. America is export-controlling atoms while China is export-subsidizing bits. One of those strategies gets stronger as models commoditize, and it is not ours.

The hardware story is more honest for the West, but the direction of travel matters more than the snapshot. Huawei's Ascend chips, the flagship of Chinese self-sufficiency, turned out on teardown to be built substantially on TSMC dies acquired through sanctions evasion, nearly three million of them routed through a shell buyer⁴. Chinese self-sufficiency in 2026 is partly theater. But look at what is happening underneath the theater. A state-backed group in Shanghai has begun low-volume production of domestic immersion DUV lithography machines, with around five shipping this year to SMIC, Hua Hong, and CXMT, and twenty planned next year. Against ASML's 131 immersion tools shipped last year⁵, that is nothing, and the Chinese machines are worse, slower, and still dependent on imported Japanese components. It is also the hardest step in the entire supply chain being taken for the first time. DUV with multi-patterning prints chips a generation or two behind the frontier, expensively, and expensive-but-domestic is exactly the kind of problem command economies are built to brute-force. On a decade view, the safe assumption is not that the moat holds. It is that the moat buys time, and the question becomes what each side does with the time.

Which brings me to the number that I think matters more than any model release, and that almost nobody in the AI debate talks about. In 2025, China added 543 gigawatts of new power capacity, an energy buildout of roughly half a trillion dollars in a single year⁶. The entire American grid added about 60. That is roughly nine to one against the country running the most power-hungry buildout in history, and China's additions since 2021 alone exceed the entire installed US grid. Meanwhile American data centers added 8.5 gigawatts of capacity last year, have 13.6 scheduled this year⁷, and demand is projected to more than double from 31 gigawatts to 66 by 2027⁸, straight into multi-year transformer lead times, sold-out turbine slots, and interconnection queues that move at the speed of litigation. Now put this next to the bottleneck migration from the last essay. The constraint on manufactured intelligence is moving from silicon, where America holds the chokepoints, toward power, where China does. Export controls are a fortress built around the bottleneck of 2024, while the bottleneck of 2030 is being cornered by the other side. If the physical constraint keeps migrating toward energy, time is not on our side, and the most important industrial policy question in America is not chip subsidies. It is why one gigawatt takes five years to connect.

Having made the case for worry, I should be clear about where I actually land, because it is not on the doom side. I am more bullish on the United States, and the reason is the substitutability rule from the last essay applied to nations instead of stocks. America's deficit is energy, and energy has workarounds. Gas behind the meter, solar and storage in the desert, siting compute where stranded power already exists, restarting nuclear plants. And the workaround list keeps growing. Elon Musk says he wants solar collected in orbit within the decade, where the sun never sets and nobody files an environmental review, and whether or not his timeline holds, the direction is the point. The menu of ways around the energy constraint gets longer every year.

Closing an energy gap is an engineering and permitting problem, ugly but solvable with money and will, and America has both, plus some of the cheapest natural gas on earth. China's deficit is the semiconductor supply chain, and that has no workaround. To close it, China has to recreate ASML, TSMC, and the hundreds of specialized suppliers beneath them, decades of accumulated process knowledge across the most intense supply chain humans have ever built, starting from five domestic DUV machines a year against 131. Both powers are racing to fix their weakness, but the weaknesses are not symmetric. America is short a commodity. China is short a miracle. That is why I believe the US semiconductor lead outweighs China's energy lead, and why the real threat to the American position is not China's grid. It is our own permitting.

Underneath the scoreboard is a more uncomfortable question. Which operating system is better suited to a buildout like this, authoritarianism or democracy? The honest answer cuts both ways. Command systems are built for pouring concrete. When the state decides power gets built, permits do not take five years, transmission lines do not die in court, and coal, solar, nuclear, and hydro get built simultaneously at whatever scale the plan demands. Democracies are built for allocating capital, and it shows. The American buildout is the largest privately financed infrastructure project in history, funded by price signals rather than decrees, and every layer of the frontier stack was invented inside open societies. But each system is now being tested at its weak point. Authoritarian buildouts misallocate on a colossal scale and answer to no market when they are wrong. Democratic buildouts underbuild and litigate, and the veto points that make democracies humane in normal times, the environmental review, the local objection, the appeal, have quietly become the binding constraint on democratic AI capacity. The race will partly be decided by which failure mode costs more this decade. And the stakes are not symmetric, because manufactured intelligence is not a neutral input to both systems. Cheap, abundant cognition is the most powerful surveillance and social-control technology ever created, which makes it a regime-strengthening technology for authoritarian states in a way it is not for open ones. The values embedded in the models the world runs on, and the system that builds the most capacity to run them, are the same question viewed from two angles.

Follow the state money and the securitization is already visible on every continent. Washington has moved from subsidy toward ownership, with \$52.7 billion of CHIPS grants, an equity stake in Intel itself, a \$500 billion public-private compute buildout in Stargate, and a Department of Energy contract to restart domestic uranium enrichment after the country outsourced essentially all of that capacity at the end of the Cold War⁹. Beijing runs the same play, a \$47.5 billion third national semiconductor fund stacked on state compute funds and a state-dominated rare earth chain. Brussels has a €43 billion Chips Act and a €200 billion AI investment program. Japan is deploying roughly \$65 billion, Korea tens of billions more, and the Gulf is building gigawatt AI campuses with sovereign wealth¹⁰. Seven of the largest American technology companies have taken nearly \$38 billion of government support over their lifetimes, SpaceX and Tesla among them⁹. Every layer of the stack, in every bloc, now has state capital behind it. That is what an arms race looks like on a balance sheet.

So where does it land? The last essay argued that on a decade view models commoditize toward the cost of tokens, with one exception. If AI becomes a first-order geopolitical concern, commoditization stops at national borders. I now think that exception is the base case at the frontier. The race dynamics push both governments the same direction. As capability starts to look militarily decisive, frontier models get treated less like software and more like enriched uranium, walled inside national security perimeters, while the commodity layer, the Kimis and Qwens and open-weight Llamas, diffuses globally and prices toward zero. Two stacks, two blocs, a commodity floor and a classified frontier. And here is the investment observation hiding in the geopolitics. Every one of those futures, commoditized or walled, cooperative or cold-war, requires more physical capacity, because a securitized race builds harder than a commercial one. Nations do not capex-discipline their way through an arms race. The bottleneck thesis is not a bet on which bloc wins. It is a bet that both keep building, and that is the safest bet in the whole series.

One consequence of an arms race in manufactured intelligence stays hidden in the national accounting, and it is what the race does to the people inside each country. The same buildout that redistributes power between nations redistributes income within them, away from wages and toward the owners of the machines, at a speed no tax system was designed for. That is the subject of the next essay.

Sources

1. DeepSeek R1 release (January 20, 2025) and Nvidia's \$589B single-day market-cap loss (January 27, 2025). Via Forbes and MIT Technology Review.
2. Kimi K3 (2.8 trillion parameters, July 2026). Moonshot AI release coverage and AI Proem.
3. Chinese open-weight share (8 of top 10) and Qwen passing Llama in downloads. Via Inference Hub, MIT Technology Review, and HuggingFace download data.
4. Huawei Ascend teardowns showing TSMC dies and the ~2.9M die Sophgo purchase. TechInsights via SemiconductorX.
5. China domestic immersion DUV production (~5 tools 2026, ~20 planned 2027, deliveries to SMIC, Hua Hong, CXMT) and ASML's 131 immersion tools shipped. Reuters reporting via Tom's Hardware and TrendForce.
6. China power additions (543 GW in 2025, ~\$500B buildout). China National Energy Administration figures via OilPrice.com and CarbonCredits.com.
7. US data center capacity additions (8.5 GW realized 2025, 13.6 GW scheduled 2026) via Goldman Sachs Research (Aterio data). US grid additions (~60 GW in 2025) and China's post-2021 additions exceeding the installed US grid via EIA and Bloomberg, January 2026.
8. US data center power demand (31 GW in 2025 to 66 GW by 2027). Goldman Sachs Research.
9. US state backing. CHIPS Act grants (\$52.7B program), the US government equity stake in Intel (Intel 8-K, August 2025), Stargate (\$500B public-private compute buildout), the DOE 10-year ~\$900M HALEU enrichment contract (General Matter, Paducah), and ~\$38B lifetime subsidies across the seven largest US tech companies incl. SpaceX and Tesla. Via the US Department of Commerce, Intel SEC filings, the US Department of Energy, and press reports.
10. China Big Fund III (\$47.5B state semiconductor fund) and rare-earth supply chain control via Caixin and SCMP. EU Chips Act (€43B) and InvestAI (€200B) via the European Commission. Japan ~\$65B chip programs incl. Rapidus (~\$18B) via METI and Bloomberg. Korea semiconductor support (\$19B+) via Chosun Daily. Gulf, the HUMAIN 1GW cluster within the 5GW US-UAE Stargate campus, via Arab News and state announcements.

The Ledger: Wealth and Work in the Machine Economy

The last essay was about how the buildout redistributes power between nations. This one is about the redistribution happening inside them, which for most people collapses into a single question, what happens to my job? The warnings are not subtle. Dario Amodei, who runs Anthropic, has said AI could eliminate half of entry-level white-collar jobs within one to five years, with unemployment reaching ten or twenty percent¹, a warning he has since begun to walk back as the data refused to cooperate. Coming from the man selling the technology, that is either admirable honesty or the best marketing in history, and either way it deserves a serious answer. Mine is more optimistic than the consensus, but not because I think the disruption is overstated. It is because I think everyone is staring at the supply side of the economy, and the answer lives on the demand side of the ledger.

First, the honest mechanics, because the disparity risk is real. The macro essay walked through what happens when capital can do cognitive work. The wage share of income falls, the capital share rises, and income flows toward whoever owns the machines. Ownership of the machines is concentrated, equity is held disproportionately by people who already have wealth, and the fastest-compounding asset in the world right now is a data center. Left entirely alone, the arithmetic of this transition concentrates income faster than any force in modern economic history. The problem is real. But notice what kind of problem it is. A distribution problem, not a production problem. The pie is exploding. The fight is over the slices, and that distinction matters because production problems make everyone poorer while distribution problems are, at least in principle, solvable.

Second, what is actually happening in the labor market in 2026, because the data is stranger than either the doom or the denial. Roughly 49,000 American layoffs through April of this year have been linked directly to AI, which against a workforce of 160 million is statistical noise². Mass firing is not happening. What is happening is quieter. Companies have stopped opening the door. Entry-level postings are shrinking, and employers are demanding experience for jobs that used to be where you got experience, a phenomenon researchers are calling experience creep³. The bottom rung of the white-collar ladder is being sawed off while the ladder itself still stands. I will not pretend to be neutral here. I am twenty-one, I have spent one summer at Silver Lake and another at Jefferies in investment banking, and I have no illusions about the outputs I was tasked with, the decks, the models, the formatting passes late at night. That work is exactly what commoditizes first, and it should, because a machine can already produce most of it. What does not commoditize is everything those jobs were supposed to teach on the way up, judgment about what the numbers actually mean, whose trust you hold, what is worth building in the first place. The problem for my generation is that the industry priced years of commodity output as the tuition for that judgment, and AI just made the tuition worthless without making the judgment any less scarce. Meanwhile, the same companies quietly closing doors to analysts are throwing them open for electricians. Ford and AT&T are ramping up recruiting for skilled trades, IBM

announced it would triple entry-level hiring⁴, and the buildout from the capex essay needs welders, linemen, and substation crews in numbers America has not trained. The signal in the noise is not disappearance. It is migration.

Third, the history, with its limit stated honestly. Two hundred years ago, four in ten American workers farmed. Today it is under two percent⁵, and the descendants of those farmers are not unemployed, they are doing jobs no farmer could have imagined. ATMs were supposed to end bank tellers and instead teller employment rose for decades as cheap branches multiplied⁶. Spreadsheets were supposed to end accountants and instead created an analysis industry. The lump of labor fallacy, the idea that there is a fixed amount of work to be divided, has been wrong every single time it has been tried, because human wants turned out to be unlimited. Every productivity gain lowered prices, freed income, and that freed income went looking for new things to want, which became new work. The honest caveat is that every previous machine complemented human cognition, and this one substitutes for it, so the pattern is not guaranteed to hold. Anyone who tells you they are certain, in either direction, is selling something.

So here is why I still land optimistic, and it is the point I think the whole debate misses. The economy does not exist to produce. It exists to consume. Every dollar of output, every token generated, every data center built, exists because a human being somewhere up the chain wants something. A cure, a house, a game, a vacation, a faster answer. The machines produce, but they do not want. A GPU never buys dinner, never upgrades to the window seat, never wants a bigger home for its family. Production can be automated. Wanting cannot. Which means that as the supply side of the economy inflates toward abundance, the binding constraint on growth stops being our ability to make things and becomes our appetite to absorb them, and that appetite is the one input only people supply. Humans do not keep their place in this economy because machines cannot do their tasks. They keep it because the entire machine economy points at them. Demand is the permanently human side of the ledger, and an economy of exploding output needs its consumers more, not less, which is precisely why every serious fiscal proposal, from compute dividends to sovereign wealth stakes, is at bottom a plan to keep purchasing power in human hands. The system does not work otherwise, and the people who own the system know it.

What does work look like on the other side? It migrates toward the two places machines point back at people. The first is the demand side itself, deciding what gets built, owning outcomes, taste, judgment, trust, relationships, the jobs where being human is the qualification rather than the limitation. When execution becomes cheap, the scarce skill is knowing what is worth executing, and the number of viable businesses explodes because the first essay's logic cuts both ways. If intelligence is rentable by the hour, the barrier to starting something falls to the price of knowing what to start. To be clear about how this sits with the macro essay, judgment is not the economy's binding constraint. Megawatts and machines are, and taste will not gate how much the machines can produce. It gates where humans fit inside that production, which is this essay's question, not that one's. More output means more firms, and every firm, however automated, has humans at the top of it deciding and consuming on its behalf. The second is the physical world.

The buildout is the largest blue-collar jobs program in American history, there are five-year backlogs on the equipment side and no queue of qualified people to install it, and the wage premium that spent forty years flowing to the college desk job is starting to flow back toward the trades. That inversion will be one of the defining social facts of the next decade.

The real danger, then, is not the existence of work. It is speed and ownership. Work migrates over years. Paychecks stop in a quarter. The wage share can fall faster than new roles emerge, and the gains are compounding in equity accounts most wage earners do not hold. Be concrete about who does hold them. The top ten percent of American households own close to ninety percent of the stock market, and more than forty percent of the country owns none of it at all⁷. The machines are publicly listed, the capital side of the ledger trades every day under tickers, and a brokerage account is the cheapest ticket in history to the ownership side of an industrial revolution, which is why I think the most consequential financial decision of my generation is whether we show up on the capital side. But a ticket still has to be paid for out of savings, and a falling wage share is precisely the thing that eats savings, so the people most exposed to this transition are the least able to hedge it. That is the disparity machine in one sentence, the same force that shrinks your paycheck inflates assets you do not own. It compounds quietly for years, then it shows up in politics all at once. This is where the underconsumption risk from the macro essay lives, and why the fiscal system will be forced to adapt whether it wants to or not, shifting the tax base from payrolls toward capital and compute, and recycling returns into demand through dividends or broader ownership mechanisms. The optimistic case in this essay is conditional on that adaptation happening. Abundance does not distribute itself.

So my read on the disparity question comes down to four claims. Abundance with a distribution problem beats scarcity with any distribution, real purchasing power rises as the price of everything cognitive collapses, humans stay load-bearing because demand cannot be automated, and work migrates to the ends of the economy machines cannot occupy, wanting and building. But the transition will be the most politically turbulent economic event since industrialization, because it is running in years instead of generations. Which leaves exactly one question this series has not faced, the one every optimistic page above quietly assumes away. That is the next essay.

Sources

1. Amodei warning (50% of entry-level white-collar jobs, 10-20% unemployment). Anthropic CEO public comments via TheStreet and Axios.
2. ~49,000 US layoffs linked to AI through April 2026. Challenger, Gray & Christmas, April 2026 report (May 7, 2026).
3. Entry-level postings squeeze and "experience creep." Via Stern Strategy Group, Washington Monthly, and Yale CELI via Fortune.
4. IBM tripling US entry-level hiring (February 2026), IBM announcement. Ford and AT&T ramping skilled-trades recruiting, CNBC, May 2026.
5. US farm employment share (~40% circa 1900 to under 2% today). USDA Economic Research Service and BLS historical statistics.
6. ATM and bank teller employment history. James Bessen, Boston University.
7. Stock ownership concentration (top 10% of households hold ~87% of equities per Fed Distributional Financial Accounts, Q1 2026, and 42% of Americans own no stock per Gallup, April 2026). Federal Reserve and Gallup.

The Wildcard: Alignment

Every essay in this series argues a version of the same claim. The buildout is real, the constraint is physical, and the value pools at the bottlenecks. An honest series has to end by naming the assumption underneath all of it, which is that the machines stay aligned with the people building them, and that nothing on that front goes wrong badly enough to stop the program. Most writing in my corner of the world either ignores the subject or waves at it on the way out. I want to do something different with it, because I think alignment, taken seriously, actually completes the thesis.

Alignment is the problem of making systems more capable than us reliably do what we intend, and it is not a fringe concern. In 2023, the chief executives of the leading labs signed a public statement that mitigating the risk of extinction from AI should be a global priority alongside pandemics and nuclear war¹. The people selling the technology signed that with their own names. Inside the buildings, alignment is discussed the way engineers discuss a known failure mode, and the doomsday view is held seriously by accredited people. A series that leans on lab revenue numbers for six essays owes their risk warnings honest treatment in the seventh.

The most common response to alignment fear is to say we should slow down. It sounds responsible, and inside a single lab it sometimes is. As national policy it fails on contact with the fifth essay, because a race does not stop when one runner does. China added 543 gigawatts of power in a single year, is standing up a domestic lithography industry tool by tool, and open-weights its best models specifically to erode the American lead². None of that pauses for a safety review, and nobody serious believes it would pause because Washington did. A unilateral American slowdown would just relocate the frontier to whichever actor is spending the least on the problem the pause was meant to solve. A pause is a handoff.

Which flips the whole question. If the race runs regardless, the choice was never race or do not race. It is lead or trail, and the two are not symmetric on safety. Alignment work costs compute, talent, and time, and those are luxuries of whoever is ahead. A trailing power cuts corners to catch up. A leading power can afford margin. To hold a model back for months of testing, to spend a real share of its compute on safety research, to say no to a deployment. The lead is the safety budget. And America's lead has a large moat because of the semiconductor supply chain, the same chain the fourth and fifth essays called the hardest bottleneck on earth. The people most worried make this exact argument. Amodei, the same chief executive whose extinction warning appears above, has argued publicly for what he calls an entente strategy, where the democracies lock down the chip supply chain, the chips, the tools, and the equipment, precisely to hold a lead wide enough that safety work can be afforded at the frontier³. The man most alarmed about the machines wants the semiconductor moat deeper. If you take alignment seriously, the chain that keeps the careful side ahead is both the scarcest economic asset in the world and the margin of safety for everyone, which makes it more valuable, not less. The caveat cuts the other way, and it

deserves its sentence. A lead only buys safety if you actually spend some of it on care, and a leader who races like a trailer has wasted the entire point of leading.

There is a second, quieter reason alignment concentrates value in the chain. Suppose the world someday gets serious about governing frontier AI. Treaties, audits, compute limits, the whole apparatus. What would enforcement even attach to? You cannot regulate an idea. Weights copy for free and leak like water. The only layer of the stack that is physical, countable, and chokepointed is the hardware, a handful of lithography suppliers, a few foundries, fabs that take years and tens of billions of dollars to build. Every compute-governance proposal worth the name runs through the chip chain, because it is the only steering wheel that exists. Race harder, and the chain is the lead. Regulate harder, and the chain is the enforcement mechanism. There is no version of taking AI seriously, optimist or doomer, that does not end with a government hand resting on the same few chokepoints this series has been describing.

There is a third development worth naming, because it changes the safety math, and it is the shift toward open source. More and more strong models are now released with their weights public, and that does two things at once. First, everyone can see how the model is built. Inside a closed lab, a few hundred employees can study the system. When the weights are open, every researcher on earth can, and the field's most important discoveries about how these models actually behave have come from exactly that kind of outside scrutiny⁴. Second, everyone has access to the same models. The defenders hold the same tools as the attackers, the small hold the same tools as the large, and nobody is at the mercy of what a closed lab chooses to sell them. The sensible caveat is the frontier, the very strongest models should be tested before they are shared, and the labs that release openly already test first. But the conclusion runs deeper than either point, and Zuckerberg has made it better than anyone⁵. The real danger was never that everyone would have this technology. It is that only a few would. Spread the models and you spread the risk with them, thinly, across billions of people with different aims, instead of stacking it inside a handful of institutions where one bad decision travels the whole world. Concentrated power is a single point of failure. Open source is how you refuse to build one. The shift toward openness is not a compromise on safety. It is a mechanism of it.

Now the honest part. I do not have an edge on whether alignment gets solved, and I distrust anyone who claims one, because the question is not answerable from public information and maybe not from private information either. A genuine alignment crisis, a failure ugly enough to force a political stop, remains the true bear case for everything in this collection, and I will not pretend it is hedgeable, because in that world every risk asset has a bad decade, not just mine. But notice what every branch short of catastrophe does. Racing makes the chain the lead. Governance makes the chain the lever. Crisis makes the chain a strategic stockpile, nationalized rather than written off. Alignment fear, the one force that reads like this thesis's enemy, puts a government floor under the hard bottlenecks in every scenario on the way to the bad one. If anything, alignment fear deepens the case for the bottlenecks. My posture is the one this

collection is named for. Conviction about the thesis, respect for the risk, watchfulness over both, and a public change of mind the moment the evidence turns.

So that is the series. The buildout is real, the constraint is physical, the value pools where supply cannot answer, the race will not pause, and the safest path through the danger runs down the same chain the money does. I know where I stand. Long the constraint, patient on the timeline, and never, ever short human wanting.

Sources

1. Statement on AI Risk, Center for AI Safety, May 2023, signed by the chief executives of OpenAI, Anthropic, and Google DeepMind, among others.
2. China power additions, domestic lithography program, and open-weight strategy. See sources in the arms race essay.
3. Dario Amodei, "Machines of Loving Grace" (October 2024), on the entente strategy and securing the chip supply chain. See also his "On DeepSeek and Export Controls" (January 2025).
4. NTIA report on open-weight models (July 2024), finding insufficient evidence of marginal harm to justify restrictions, and interpretability research built on open weights, incl. Arditi et al., Refusal in Language Models Is Mediated by a Single Direction (2024).
5. Mark Zuckerberg, "Open Source AI Is the Path Forward" (July 2024) and his August 2026 essay on the concentration of power.

The Future of Tokens: Demand at Machine Speed

The macro essay argued that human labor stops being the binding constraint on output once capital can do cognitive work. This one is about what that does to the meter. Every argument in this collection eventually reduces to a single number, which is how many tokens the world is going to need, and nearly every forecast of that number is built on a picture of a person typing. That picture is already obsolete. Human-scale usage, one prompt and one answer, was never the market. It was the demo.

Start with what is actually being counted, because the unit is easy to wave past. A token is a piece of a word. Models read and write in these pieces, and every one of them is metered and billed, so a token is both the atom of machine thinking and the line item on the invoice. The Declaration of Independence is about 1,600 of them¹. A long conversation with a model might run a few thousand. Those numbers are why token counts sound small and harmless, and they are also why the market keeps underestimating them, because the workloads that matter no longer look anything like a document or a conversation.

Agentic work runs in loops. Plan, act, observe, verify, retry. A model asked a question answers once. An agent given a job writes a plan, calls tools, reads what came back, checks its own work, finds the error, and goes again, sometimes for hours, and every pass through that loop is inference. Gartner puts the gap at five to thirty times the tokens of a comparable chat request, and Stanford researchers measuring coding agents against simple code chat found it running to a thousand times². Priced out, one completed agentic task runs between ten cents and a dollar against a fraction of a cent for the chat exchange it replaces². To anyone who does not work with these systems every day, asking a model a question and handing an agent a job can sound like the same request. It is not the same work. The agent is doing far more, and it is consuming orders of magnitude more to do it.

Then there is the clock, and this is the part I think is most underpriced. A human-driven system consumes tokens at the speed of human attention, because someone has to read the output, decide, and prompt again. That caps demand at waking hours and the pace of one person's judgment. The third essay argued that human cognitive labor stops being the constraint on what the economy can produce. Read that claim off the meter instead of the ledger and it says something sharper. Take the person out of the middle of the loop and the ceiling on consumption leaves with them. An agent that verifies its own work does not stop at five, does not sleep, and does not wait for a colleague to approve step four before it starts step five. Every hour the review step used to take becomes an hour of inference instead. The economy has spent its entire history metering cognitive work by the availability of the people doing it. That meter is coming off.

The second-order effect is larger than the first. As agents begin transacting with other companies' agents, scoping work, negotiating terms, dispatching jobs, confirming completion, every commercial interaction that used to be a phone call or a flat file becomes a multi-turn exchange

between two reasoning systems, each burning inference on both sides of the trade. This is not speculative. The plumbing is being laid right now. The protocol that lets agents call tools has been downloaded 97 million times and adopted by every major lab, the protocol that lets agents discover and negotiate with each other has more than 150 participating organizations and sits at the Linux Foundation, and there is now a payments standard whose entire purpose is proving that a human authorized what an agent just bought³. There is also precedent for how large machine-initiated volume gets once rails like that exist. Visa has found that only about a tenth of the transactions crossing its network are initiated by a person touching something⁴. The rest is already software talking to software. Commerce made that transition once, quietly, on rails built for something else. It is about to make it again, except this time both ends of the conversation are thinking.

And the loop feeds itself. Agents produce outcomes, and outcomes are training data of a kind the internet never contained, a record of what was actually tried and what actually worked. The more work the economy routes through agents, the more valuable the next generation of models becomes and the more compute it takes to build. Inference demand recycles into training demand. Notice how completely that inverts the old arrangement. For fifty years we trained people to use software. Now the software is the thing being trained, and every hour it runs is another hour of instruction. That turns the usage itself into an asset rather than a cost. A machine wears out as you run it. This one gets better, because every job an agent finishes is another example the next model learns from, and a better model gets handed more work, which produces more examples again. It is also why the familiar technology curve does not apply. A maturing product normally needs less compute per unit of output every year, and this one takes every efficiency it earns and spends it on a bigger job.

The obvious objection is price. Inference costs have fallen roughly a thousandfold in three years⁵, so a skeptic can argue that all this volume arrives at a collapsing unit price and the revenue never shows up. The macro essay named why that argument fails, because when the price of a valuable input collapses, demand does not politely stay where it was. There is a sharper version here. Falling token prices are not a headwind to this thesis, they are the mechanism of it. Agentic loops are only economic because tokens got cheap. At 2023 prices, spending a thousand steps of verification to finish one task would have been absurd, and at today's prices it is a rounding error against the salary of whoever used to do that task. Every price cut converts another category of work from too expensive to automate into obviously worth automating, and the volume that walks through the door is larger than the price decline that opened it.

And the largest source of that volume is not the work agents take from people. It is the work agents make that nobody could afford to do before. Replacing existing work comes first, because it is the easiest kind of spending to approve. A company already knows what that work costs it today, so the savings are simple to point at. The volume that follows comes from everything that was never done at all. The analysis a small business could never buy. The second opinion

nobody had time to seek. The thousand molecules never screened because screening them cost more than the answer was worth. Every one of those is a job that did not exist as demand last year and is a metered workload this year. When something scarce becomes abundant, people do not buy the same amount of it at a lower price. They start buying things they could never afford to buy at all.

Put all of that together and you get a decoupling I think is the most underpriced fact in this collection. Token consumption stops scaling with the number of humans typing and starts scaling with the volume of economic activity itself. Those are different curves with different ceilings. The first is bounded by population and attention and grows a few percent a year at best. The second is bounded by GDP. Which means compute, and the power underneath it, stops being a line item in an information technology budget and becomes a direct input cost to output, closer to electricity or freight than to software licensing. Consensus is still modeling the first curve. Google was already processing more than three quadrillion tokens a month by last spring, seven times the prior year, before agents were anything more than a pilot in most companies⁶, and the forecasts that try to price the agentic era have consumption multiplying more than twenty times again by 2030⁷.

So this is not a software cycle. Software cycles are gated by budgets and seat counts, they saturate, and the marginal copy is free. This is an infrastructure buildout with a demand curve tied to the whole economy, and every unit of it is paid for in silicon, electricity, and cooling. It is capital constrained at every layer, generation, interconnection, silicon, and the research that keeps the loop compounding. Which leaves one question, and it is the one the final essay has to answer. If the demand is this visible and the constraints this physical, why is almost none of it in the price?

Sources

1. Token counts for reference documents (Declaration of Independence ~1,627 tokens): Ribbit Capital, "Token Letter," June 2025; OpenAI tokenizer documentation.
2. Agentic token multipliers (Gartner: 5-30x tokens per task vs. chatbots, March 2026; Stanford coding-agent measurements up to 1,000x vs. simple code chat; \$0.10-\$1.00 per completed agentic task vs. fractions of a cent per chat call; 50,000 tokens per code review, 200,000 per browser procurement task): Gartner via Spheron; AgentMarketCap; Zylos Research.
3. Agent protocol adoption (Model Context Protocol ~97M downloads, adopted by Anthropic, OpenAI, Google, and Microsoft; Agent2Agent protocol with 150+ participating organizations, donated to the Linux Foundation June 2025, v1.0 with signed agent cards in early 2026; Agent Payments Protocol for cryptographically verifiable purchase authorization): Linux Foundation; Google; Cipher Projects; Digital Applied.
4. Visa analysis finding roughly 10% of network transaction volume is human-initiated: Visa, "Making Sense of Stablecoins," April 2024.
5. Inference cost decline of roughly 1,000x over three years: Ribbit Capital, "Token Letter," June 2025; Epoch AI.
6. Google processing 3.2 quadrillion tokens per month, seven times year over year: Google I/O 2026.
7. Forecast of token consumption multiplying ~24x to ~120 quadrillion tokens per month between 2026 and 2030: Zylos Research; industry forecasts compiled by AgentMarketCap.

The Opportunity: Crowded Theme, Empty Trade

Anyone who has read this far should now raise the obvious objection. If the buildout is the largest in the history of capitalism and the bottlenecks are visible enough that a college student can list them, then surely it is all in the price. And by every measure of positioning, it should be. In Bank of America's August survey of global fund managers, long the Magnificent 7 was once again named the most crowded trade in the world¹. Two months earlier the same survey found 80% of managers crowding into global semiconductors, one of the highest conviction readings in its history². Hedge funds just raised their tilt toward technology by the most in any quarter on record³. These are the most owned, most covered, most discussed companies on earth. I am not bringing anyone a new ticker. So the fair question, the one I would ask any manager pitching this thesis, is simple. Where exactly is the opportunity in the most crowded trade in the world?

Here is my answer, and it is the whole essay. The macro has believers. The micro has none. Money is never static, it flows, and you learn more from where it refuses to go than from where it goes. At the level of the theme, the survey level, the index level, belief is total. Then walk down the ladder toward the actual securities, the actual order books, the actual signatures, and watch the belief drain out at every rung. A survey measures what people say. A price measures what they pay. Ignore the commentary and read the tape. The tape says the market that cannot stop talking about AI is pricing the companies that supply it as if the whole thing ends in about two years.

Start at the top rung, the biggest companies in the world, where the doubt is already visible. Alphabet, which runs the largest custom silicon fleet outside Nvidia's and just guided to as much as \$190 billion of capex, trades near 16 times earnings while the index trades above 20⁴. A discount to the market that contains it. Amazon raised its 2026 capex to \$220 billion, twenty of those billions because memory got more expensive, and its stock has been pinned in a range all year while investors debate whether the spend will earn its cost of capital⁵. And when these companies report, the market grades them on one axis. In July, Meta guided to as much as \$145 billion of capex and fell 7%. Microsoft guided to more, \$175 billion, and rose 7%, because it could point to Azure crossing \$100 billion of annualized revenue, a visible line from spend to sales⁶. Oracle raised capital spending 162%, announced \$40 billion of new debt and equity, and fell 12% in a day. It has been nearly cut in half since June and was downgraded to one notch above junk while sitting on a \$638 billion backlog, eleven years of its current revenue, already signed⁷. One money manager summarized the regime in a sentence. Capex used to be the more the better. Now it is the less the better⁶.

Hold that up against 1999 for a second, because the contrast is the tell. In a bubble, the market pays for the story. Announcing fiber capex made a telecom stock go up, and the checks chased the narrative all the way to the grave. In 2026, announcing AI capex makes your stock go down unless the revenue is already on the tape. The market is rewarding harvest and punishing

planting. That is not a mania. That is a market that owns the theme and does not believe the numbers.

One rung down sits Nvidia, the most important company in the world, trading at roughly 24 times forward earnings, its cheapest multiple in years and less than half its own three-year average of 52⁸. The S&P 500 trades at 20⁹. Four turns above the index. That is the entire premium, while the current quarter is guided to \$91 billion of revenue, a billion dollars a day, with the data center segment growing 92%, and consensus already near \$390 billion for next year¹⁰. Pay four turns over the market multiple, receive growth running roughly ten times the market's. Euphoria was Cisco at over 100 times earnings in March 2000. This is the opposite creature. This is the market renting the most analyzed stock on earth one quarter at a time because it does not trust the earnings past next year.

Another rung down, the suppliers, and now the doubt turns to open disbelief. Micron has sold out substantially all of its DRAM and high-bandwidth memory output through 2027 and signed sixteen multi-year customer agreements¹¹. The stock is up over 200% this year and trades at 5.7 times forward earnings against a ten-year average of 22. SK Hynix, the HBM leader, trades at 4.9 times after rising roughly ninefold. Samsung sits near 7¹². Super Micro booked over \$60 billion of new orders in a single quarter and trades at 17 times trailing¹³. And the disbelief is not a mood, it is published. Bank of America's base case for 2028 assumes DRAM prices fall 10% and NAND falls 18%, and Micron still earns roughly \$150 a share. Its severe case, a full traditional memory bust, still earns about \$100, eight times the peak of the entire 2018 cycle¹⁴. The stock trades at five and a half times the first number and about eight times the second. Say that in plain English. Micron sells about a fifth of the world's DRAM¹². Buy the stock today and the company earns back its entire share price in about five and a half years in the base case, with memory prices falling the whole way. In a full bust it takes about eight years. The downside case is a slower payback, not a loss. The market is not pricing a downturn. It is pricing the memory of every downturn since the 1990s onto a business whose worst case now out-earns its best pre-AI year by a factor of eight. Meanwhile the buyers are telling you the direction of travel. Amazon just raised its capex guide by \$20 billion and blamed memory prices⁵. A multiple of 5 is not what belief looks like. It is a countdown.

Widen the lens past the names everyone argues about, because the same mispricing runs down every aisle of the store. The equipment makers, Applied Materials, Lam Research, and KLA, the companies that sell the machines every fab on earth is built from, just raised their own 2026 spending forecast by 30%, to as much as \$160 billion, on a consensus path toward roughly \$260 billion by 2028. The stocks sold off 20 to 40% anyway, and the group now trades near 0.7 times its expected growth¹⁵. Storage is the least glamorous aisle in the data center, and it is gone too. Western Digital has sold 100% of its 2026 hard drive production, with purchase agreements reaching into 2028 and 2029, and Seagate's data center capacity is allocated through 2027¹⁶. Sold out for years is now the norm all the way down to the spinning disk. So is the discount.

Go inside the machine itself, because the chips are changing and the market has not repriced the change. Walk the chain. Nvidia alone will sell roughly \$390 billion of AI chips next year¹⁰. Custom chips, the ASICs that Google, Meta, Microsoft, OpenAI, and Anthropic design for their own data centers, are growing three times faster than GPUs and pass them in units shipped next year. Two companies, Broadcom and Marvell, design 95% of those custom chips¹⁷. Broadcom's slice alone is guided to \$100 billion of AI revenue in 2027, roughly double its entire company today, and the stock costs about 22 times the earnings that revenue implies, the same multiple as an average S&P company¹⁸. A duopoly on the fastest-growing chip category on earth, priced like everything else. The pure inference challengers get the opposite treatment. Cerebras came public at \$48 billion and trades north of 50 times sales¹⁹. The one I watch closest is Qualcomm, because it is not selling a card, it is selling full server racks, shipping this year, built by the company that spent twenty years making phone chips where every watt and every degree of heat mattered, now aimed at data centers that are rationed by exactly those two things²⁰. And here is the detail that ties every one of these machines back to the anchor. Open any of them up. Qualcomm's card carries 768 gigabytes of memory. Nvidia's next flagship carries up to a terabyte per GPU. The compute die is a sliver of the board, and the rest is memory, because serving a model means holding the model and its context right next to the processor²⁰. That is why I do not need to pick the winner of the architecture war. Every one of these designs, whoever builds it, is mostly memory, and the companies selling that memory trade at 5 times earnings.

The bottom rung is CoreWeave, the purest public expression of the buildout, and this is where the gap stops being a discount and becomes an absurdity. Watch the order book move in real time. Signed contracts stood at \$60.7 billion at the end of last year, \$99.4 billion in March, \$104 billion in June, and \$129 billion as of this week, counting the more than \$25 billion of new commitments signed in the first six weeks of the third quarter alone²¹. That last stretch is over \$600 million of new signatures a day. Roughly 40% converts to revenue within two years and about 80% within four, and the business underneath is compounding on schedule, with revenue up 112% last quarter and guided to roughly double for the year²¹. Contracted power reached 4.2 gigawatts, two Hoover Dams of electricity, secured in a country where a new grid connection takes five or more years²². The queue is the moat. A contracted watt is a call option on every GPU generation of the next decade, and the strike was paid when the substation was signed.

Now the napkin, in three steps. Step one, what a megawatt costs. Building an AI-ready data center runs \$15 to \$25 million per megawatt before the chips go in, and \$30 to \$45 million with them²³. Step two, what the market charges for CoreWeave's megawatts. The company's enterprise value, meaning the stock plus every dollar of its debt, is about \$84 billion, and its contracted power is 4.2 gigawatts, which is 4,200 megawatts. Divide one by the other and the market's price is about \$20 million per megawatt²². Step three, compare. The market is selling CoreWeave's power pipeline for roughly what the empty buildings cost to construct, which means the \$129 billion of signed contracts, the installed GPU fleet, and the operating business are all priced at zero. The bear case stays on the page. OpenAI is roughly \$22 billion of the

commitments, the net loss widened to \$626 million as the capex ramp ate the income statement²¹, and building the contracted capacity means spending about five times trailing revenue this year²⁴. So the position is sized like what it is, a leveraged claim on signatures whose largest counterparty burns cash. But the stock jumped 14% on these numbers and still sits roughly a third below its high, which is a preview of the mechanics. When the signatures print, the repricing runs into this trade, not out of it. And it is not one company. Nebius carries close to \$50 billion of signed contracts from Microsoft and Meta against roughly half a billion dollars of last year's revenue, and took a \$2 billion equity check from Nvidia²⁵.

Then follow the wire out of the building, because the same mispricing is sitting in the electrons. Power is the one input everyone concedes is scarce, and the honest prices prove it. PJM, the largest grid market in America, just cleared its capacity auction at a record \$329 per megawatt-day, roughly ten times the level of two years ago, pinned against a regulatory cap that kept it from going higher, with data centers driving 40% of the bill²⁶. Microsoft signed twenty years at an estimated \$110 to \$115 per megawatt hour to restart Three Mile Island, and hyperscaler nuclear deals now clear at roughly double wholesale power²⁷. The auction and the contract are the two honest marks for a firm watt, and both have exploded. The equities have not. Vistra, holding more than 5,000 megawatts of hyperscaler agreements including twenty-year Meta nuclear deals, trades at 16 times forward earnings with growth compounding at 37%, a PEG of one half. Talen, whose Susquehanna plant carries a seventeen-year, \$18 billion Amazon contract, trades at about 18 times forward earnings while growing them more than 30% a year²⁸. And the exception proves the rule of this whole essay. GE Vernova, which sells the turbines and books the shortage as revenue today, trades at 63 times earnings after a 243% year²⁹. The market pays the seller of shovels 63 times and the owner of the mine 16. Twenty-year signatures from investment-grade counterparties are the cheapest signatures on the tape.

And underneath the whole ladder, the belief goes negative outright. Short interest against US equities just hit a record \$2.13 trillion. The median S&P 500 stock has not been shorted this heavily since 2011, and total NYSE short interest now exceeds its financial crisis and pandemic peaks³⁰. Nearly half of the managers in that same BofA survey call an AI bubble the market's biggest tail risk, roughly double the share of a month earlier³¹. Put the full tape together. Owned at the index, doubted at the multiple, shorted at the security. The theme is crowded. The belief, walking down rung by rung, thins, sours, and at the bottom flips into an active bet against.

So take the market's side, because it deserves its strongest case. Maybe the earnings really are a peak and a single-digit multiple is fair value on a dying cycle. There is a way to check, and it is the check power traders run every morning. In electricity markets, the spark spread is the gap between the price of power and the cost of the fuel that makes it. As long as the spread is positive, you run the plant. Compute has a spread too, and the best place to read it is the oldest chip still working. The factories essay used the H100 to argue that old chips stay alive. Here is the same machine as a trade. It is four years old now, two full generations behind the frontier. At the peak of the 2023 shortage it rented for \$8 an hour. By last October it had slid to \$1.70, and

the depreciation bears looked right on schedule. Since then it has climbed back to \$2.35 on one-year contracts, up almost 40%, because on-demand capacity is sold out across the market³². Now the cost side of the spread. The chip sells for \$30,000 to \$40,000. Spread the middle of that range over four years of round-the-clock work and it costs about a dollar an hour to own, call it a dollar and a half all-in with power and facility. Renting at \$2.35 against a dollar and a half of cost is a positive spread on a machine the accounting says should be nearly worthless, and the spread is widening. And that is only the first spread. The second sits on top, where two dollars of machine time substitutes for fifty dollars of knowledge work. Spreads like that die exactly one way, through a supply glut. So show me the glut. ASML ships EUV machines in the dozens per year. Advanced packaging is sold out into 2027. Memory is sold out through 2027. TSMC, the one company that sees every real order book on earth, just raised its 2026 growth outlook above 40% and lifted its own capex to \$64 billion³³. The fuel is rationed, the power price is rising, and the market is pricing the plants for the end of demand. And the spread steepens with every generation. CoreWeave's published benchmark of Nvidia's next platform runs ten times the tokens per megawatt of the current one, so the same watt earns more rent with every refresh³⁴. There is an old version of this setup. In the 1860s the money in oil was never in the wells, it was in refining, and the market of that era needed a decade to notice. The refiners of intelligence are sitting in plain sight at five times earnings. The market owns them. It just refuses to believe the spread.

Now the shape of the bet, because the asymmetry is the point of the whole structure. The downside is a multiple that already assumes the earnings die, on order books that run two years past the assumed funeral. The upside is what happens when cycle multiples meet regime earnings, and it is not capped by anything except the size of the buildout. Known downside, uncapped upside. Crowded trades are dangerous when the crowd has priced perfection and has to run for one exit at once. This crowd has priced the funeral. There is nothing to give back, and the repricing, when it comes, runs into the trade rather than out of it. You do not have to imagine what that repricing looks like, because it has already happened in exactly one aisle. NAND prices are forecast to rise 234% this year with the shortage running into 2028, and SanDisk, the purest way to own that corner, rose 726% in the first half of the year, the best stock in the S&P 500³⁵. One aisle repriced, and it repriced violently. And even the repricing was disbelief in disguise. SanDisk's earnings grew twenty-two fold while the stock rose sevenfold, which means the best performer in the index got cheaper as it climbed. The rest of the store is still marked for clearance. My edge is not information. Every number in this essay is public and most of them sit on free websites. The edge is structure and nerve. Underwrite the magnitude consensus refuses to, in the layers where the signatures already prove it, and hold without leverage while the market argues with its own tape.

Strong beliefs deserve tripwires, so here is what would prove the market right and me wrong. HBM contract prices cracking while capacity is still being added. CXMT, China's memory champion, reaching competitive yield at scale years ahead of schedule. Rental rates rolling over

instead of rising. Backlogs and remaining performance obligations stalling at the neoclouds, or an anchor counterparty failing to pay. Hyperscaler power contracts and capacity auctions repricing lower at renewal. Token growth flattening while ASML's shipment count ramps. Those are the signatures of a real glut forming, and the day the order books stop growing before the multiples re-rate, the cycle case wins and I will say so in writing. Nothing on the tape says that today.

That is the collection. The machine is real, the constraint is physical, the value pools at the bottlenecks, and the market, for all its crowding, still prices the constraint like a cycle. The gap closes one way or the other. Either the bust arrives on schedule and these essays are wrong, or the earnings keep arriving and the multiple has nowhere left to go but up. I know which side the signatures are on. The only question left is not about the machine at all. It is about who is holding the brush.

Sources

1. BofA Global Fund Manager Survey, August 2026 ("Long Magnificent 7" most crowded trade, cited by 45% of 169 managers, \$413B AUM): Reuters; Investing.com.
2. BofA Global Fund Manager Survey, June 2026 (80% of managers long global semiconductors, third-highest conviction reading on record): Benzinga.
3. Goldman Sachs Q2 2026 hedge fund positioning data (largest quarterly increase in net Information Technology tilt on record): Goldman Sachs Prime Services via press reports.
4. Alphabet at ~16x earnings vs. S&P 500 above 20; 2026 capex guidance of \$180-190B: Yahoo Finance; 24/7 Wall St.
5. Amazon 2026 capex raised from \$200B to \$220B on higher memory costs; forward P/E ~29; EV/EBITDA ~14: 24/7 Wall St; App Economy Insights.
6. Meta \$130-145B capex guidance and -7% reaction, ~29% off its highs; Microsoft \$175B guidance and +7% on Azure crossing \$100B of annualized revenue; "less the better" quote (Jason Lemire, Bold Wealth Partners): Fortune; 24/7 Wall St; FinanceFeeds.
7. Oracle capex up 162% to \$55.7B, \$40B of new debt and equity, one-day -12%, ~47% decline since June, S&P downgrade to BBB-, \$638B backlog: Reuters; The Motley Fool; EBC Financial Group.
8. Nvidia forward P/E ~24.5 vs. its three-year average of 52.4 and five-year average of 60.5: GuruFocus; The Motley Fool.
9. S&P 500 forward P/E ~20 (August 2026): FactSet Earnings Insight.
10. Nvidia Q2 FY2027 guidance (\$91B revenue, data center +92%); consensus FY2027 revenue ~\$391B: S&P Global Market Intelligence; Simply Wall St.
11. Micron DRAM and HBM output sold out through 2027; sixteen multi-year customer agreements: company commentary via TS2 and MarketWise.
12. Micron at 5.7x forward earnings (ten-year average 22, up 214% YTD); SK Hynix at 4.9x forward after a ~9x move; Samsung near 7x forward; global DRAM revenue share Q1 2026 (Samsung 38%, SK Hynix 29%, Micron 22%): The Motley Fool; GuruFocus; Investing.com; Korea Times.
13. Super Micro \$60B+ of new orders in fiscal Q4, record backlog, ~17x trailing earnings: 24/7 Wall St; Yahoo Finance.
14. Bank of America 2028 memory scenarios (base case of DRAM -10% and NAND -18% with ~\$150 Micron EPS; severe downturn ~\$100 EPS, roughly 8x the 2018 cycle peak): BofA Global Research via Benzinga and 24/7 Wall St.
15. Semicap 2026 WFE guidance raised ~30% to \$150-160B; consensus \$156B/\$213B/\$262B for 2026-28; sector off 20-40% at ~0.7x three-year PEG: BofA via Investing.com; Morningstar; Morgan Stanley.
16. Western Digital 100% of 2026 HDD production sold out with agreements into 2028-2029; Seagate nearline capacity allocated through calendar 2027: Tom's Hardware; Trefis; Yahoo Finance.
17. Custom ASIC shipments tripling 2024-2027, surpassing GPU shipments by 2027; Broadcom and Marvell at ~95% of co-design: TrendForce via TechTimes; Tom's Hardware; Bloomberg Intelligence.
18. Broadcom AI revenue +143% to \$10.8B in Q2 FY26; \$100B AI revenue target for 2027; ~22x on FY27 estimates: Broadcom via SEC; The Motley Fool; TIKR.
19. Cerebras IPO at ~\$48B (May 2026, \$5.5B raised, +68% day one); subsequent ~28% decline on margin guidance; FY26 core revenue guide ~\$860M: TechCrunch; CNBC; Morningstar.
20. Nvidia licensing Groq technology (December 2025); Qualcomm AI200/AI250 rack-scale inference servers (768GB LPDDR per card, mobile-derived power efficiency, AI200 shipping 2026); Nvidia Rubin Ultra HBM capacity up to ~1TB per GPU: Data Center Knowledge; NAND Research; Spheron; Tom's Hardware.

21. CoreWeave Q2 2026 results (revenue \$2.58B, +112%; revenue backlog \$60.7B at Dec 31, \$99.4B at Mar 31, \$104.2B at Jun 30, +246% YoY, \$129.2B as of Aug 11 including \$25B+ of Q3 commitments; net loss \$626M; FY26 guide \$12.4-13.2B revenue; shares +14% after hours; ~40%/~80% backlog conversion within 2/4 years): CoreWeave Q2 2026 earnings release (SEC); CNBC, Aug 11, 2026; Investing.com; TradingView.
22. CoreWeave contracted power expanded from 3.7 to 4.2 GW (Q2 2026 call); market capitalization ~\$50B and enterprise value ~\$84B post-report: CoreWeave earnings call via 24/7 Wall St; GuruFocus; StockAnalysis.
23. AI data center build costs (AI-optimized shell and infrastructure \$15-25M/MW excluding GPUs; \$30-45M/MW fully loaded): Axis Intelligence; JLL benchmarks.
24. CoreWeave 2026 capex guidance of \$31-35B against ~\$6.2B of trailing revenue; stock roughly a third below its 52-week high post-report: The Motley Fool.
25. Nebius: Microsoft contract up to \$19.4B, Meta up to \$27B, ~\$50B total contracted backlog vs. \$530M 2025 revenue; \$2B Nvidia equity investment: CNBC; The Motley Fool; TradingKey.
26. PJM 2026/27 capacity auction record \$329.17/MW-day, ~10x in two years, price-capped; data centers ~40% of the \$16.4B cost: Utility Dive; IEEFA; Citizens Utility Board.
27. Hyperscaler nuclear PPAs at \$100-140/MWh on 20-year tenors; Microsoft-Constellation Three Mile Island restart at ~\$110-115/MWh (Jefferies estimate), 20-year, ~\$16B: Jefferies via AOL; industry reports.
28. Vistra (5,000+ MW hyperscaler PPAs incl. ~2,600 MW 20-year Meta nuclear deals; ~37% forecast earnings growth; ~16x forward P/E, ~0.5 PEG); Talen (17-year, ~1.9GW, ~\$18B AWS agreement; ~18.6x forward P/E with ~32% forecast annual earnings growth): TIKR; The Motley Fool; Simply Wall St; GuruFocus; Utility Dive.
29. GE Vernova at ~63x next-twelve-month earnings after a +243% one-year run; record 116 GW gas backlog: TIKR; The Motley Fool.
30. Record \$2.13T of short interest against US equities; S&P 500 short interest ~3.79% of float, highest since 2010; NYSE ~9%, above financial crisis and pandemic peaks; median S&P 500 stock most shorted since 2011: Bloomberg; S&P Global Market Intelligence.
31. Roughly half of BofA survey managers naming an AI bubble the biggest tail risk, up from 28% the prior month: BofA Global Fund Manager Survey via Bloomberg.
32. H100 rental history (\$8/hr peak 2023; \$1.70/hr October 2025; \$2.35/hr one-year contract March 2026, up ~40%; on-demand sold out); purchase price \$30-40K: GPUSmith; IntuitionLabs; CloudZero.
33. TSMC 2026 revenue growth outlook raised above 40%, capex lifted to \$64B, HPC/AI at 61% of revenue: TSMC earnings commentary via Yahoo Finance, Investing.com, CNBC.
34. CoreWeave-published Vera Rubin benchmark at 10x tokens per second per megawatt vs. Grace Blackwell: NVIDIA; Introl.
35. Gartner forecast of NAND prices rising 234% in 2026 with shortage into 2028; SanDisk up 726% in H1 2026 (best S&P 500 performer) on ~22x EPS growth: The Motley Fool; TheStreet; Forbes.