

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

Shay O'Kelly · August 2026

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: $gY = gA + \alpha \cdot gK + (1-\alpha) \cdot gL$. 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.

No Offer or Solicitation. The only purpose of this site is to provide general background information on RAXIA. This is not an offer to sell or the solicitation of any offer to buy any securities. Any such offer would be made only to qualified investors after distribution and return of the Private Placement Memorandum and governing agreement for the applicable Fund (the "Offering Materials"). Nothing on this site is investment advice or a recommendation of any security. The author may hold positions in companies discussed. Past performance is not indicative of future results.