

ESSAY · ENTERPRISE AI

The AI Capacity Fork — More Throughput or a Different Team Shape

If AI creates capacity there are two honest destinations for it, and most organisations pick neither. On reconfiguring the unit of delivery, and cutting elastic capacity in the right order.

AI STRATEGY

LEADERSHIP

TRANSFORMATION



Burak Galiba · August 6, 2026 · 6 min read
<https://burakgaliba.com/blog/the-ai-capacity-fork>

SCAN TO READ ONLINE

When Uber capped agentic coding spend at \$1,500 per engineer in June, after [burning its annual AI budget in four months](#), the argument in public was about the number. It usually is.

The number is the smaller question. An organisation that buys capacity and never decides what the capacity is *for* will have the cost argument every year, and win it every year, and still have nothing to show at the end.

The fork nobody says out loud

If AI genuinely creates capacity, there are two honest destinations for it.

Increase throughput — deliver more, faster; burn down the backlog; ship the things that have been queued for two years.

Or reconfigure the unit of delivery — change the shape of the team so the same work is done by a smaller AI-capable cell. Two people and AI owning a product outright, where the conventional shape was a squad. New work then gets staffed out of the cells that already work that way, rather than by growing teams that don't.

That's the list. Everything else is capacity going nowhere.

The speed in that second option doesn't come from the tokens. It comes from collapsing the distance between deciding what to build and building it — no ticket, no refinement, no handoff, no translation between someone who knows what is wanted and someone who knows what is possible. AI is what makes the smaller shape survivable. Removing the handoffs is what makes it fast.

Most organisations pick neither, and picking neither is rarely a decision anybody makes in a room. Licences get bought, the cost gets absorbed, and nothing structural changes: same team shapes, same roadmap, same intake, same approval chain. The capacity has no destination, so it dissipates, and twelve months later the invoice is the only durable artifact of the exercise.

EY's fourth US AI Pulse Survey — 500 senior US decision-makers, fielded September to October 2025 — found 96% reporting productivity gains, then asked what they did with them: [47% reinvested in existing AI capabilities](#), [42% in new ones](#), 29% in price cuts, 17% in reduced headcount. Some of that is real capital formation — a platform built now to serve a hundred use cases later is investment, not waste. But the two most common destinations for AI capacity are more AI, which defers the moment of return rather than closing the loop, with no date attached.

Reconfiguration is the easiest thing here to fake

"We're reconfiguring, not cutting" is the single best way to change nothing while sounding decisive. So make it falsifiable. **Reconfiguration is real if and only if one of two numbers moves:** the intake — more delivered per period, against a baseline you wrote down first — or the money you spend on capacity you don't employ, together with the hiring plan. If neither moves, the org chart got redrawn and the bill stayed the same.

The arithmetic has to stay honest too. If two people and AI do what six did, and the queue of work isn't growing, the organisation needs fewer people. "Redeploy" is an honest word only when there is somewhere real and funded to redeploy to; when there isn't, redeploy means we will not backfill — and it's better to say that sentence than to let people find it out.

Which capacity goes first is where I think most organisations have it backwards.

Outsourced and contracted development is bought as elastic capacity. That is its entire purpose: it's what you buy because permanent capacity can't flex on the timescale the work demands. AI is elastic capacity as well — instant, scaling with demand, costing nothing when idle. As substitutes for marginal delivery capacity, not for everything a supplier brings, the two do the same job. Add one and the first line that should shrink is the other.

Most do the reverse. The external contract is a signed number with a term, a supplier relationship and a renegotiation cost, so it gets protected. Internal headcount is where the org chart is, so that's where scrutiny lands. The committed spend survives; the permanent capability absorbs the pressure. That is precisely the wrong way round — and note that the contractors were usually issued licences too, so the organisation is now paying twice for the same elasticity.

The renewal date is the honest place to start, because it's the only one that arrives on its own.

The counter-argument deserves its due

There is a serious objection to all of this, and it isn't "AI doesn't work". It's that throughput isn't yours to choose. Speed up code generation in a system where the binding constraint is review, testing, release approval or decision latency, and you haven't increased throughput — you've lengthened the queue in front of the constraint.

Google's [2025 DORA report](#), covering nearly 5,000 technology professionals, found AI adoption associated with higher delivery throughput *and* higher instability — more change failures, more rework. Its framing of review is exact: velocity gains for an author become significantly increased cognitive load for the reviewer, who is still expected to audit every line. Its headline conclusion is the one for the slide: AI doesn't fix a team, it amplifies what's already there.

That objection lands hardest on the small cell, which is where it should. A cell with one developer has no second pair of eyes, no on-call rotation, no cover for leave, and no separation between the person who writes a change and the person who approves it — and AI can hold neither end of that, because it cannot be accountable for a production decision. [LinearB's platform data](#) — vendor telemetry, not research — has AI-assisted pull requests running about 2.6 times larger than unassisted ones, and agentic ones idling over five times longer before a reviewer picks them up. A cell can make the downstream constraint worse while feeling faster from the inside.

Which is why composition matters more than size. Two people who can both ship — a product owner who reads diffs, or two developers who think about the product — gives you a bus factor of two, a rotation, and an author who isn't also the approver. What it doesn't give you is independence: a fixed pair reviewing only each other converges on shared blind spots within months. So the cell owns its

routine work, and anything over a risk line — schema changes, auth, payments, anything with a compliance control attached — goes to a reviewer outside it. That threshold is a structural change downstream, which is the whole thesis: capacity doesn't convert until you move the constraint, and moving it is the change-management work that is the project rather than the postscript.

Name the fork per team, before the licences land

Throughput or reconfiguration, with the number that should move written next to it. If it's throughput, name the constraint the capacity is supposed to relieve and the delivery number that should change — and accept that something downstream has to change for it to move at all. If it's reconfiguration, say which shape, which work, and what happens to the people whose team stops existing in that form.

What you cannot do is buy the capacity, decline to name a destination, and then treat the bill as a surprise. That isn't neutrality — it's the permissionless spending an operating model exists to prevent.

And if neither destination is acceptable — nothing more delivered, nothing changed about how teams are shaped — then the correct decision is not to buy the licences. That's a defensible position. It just isn't the one most organisations think they're taking when they take it.

If you want a read on whether your organisation has the machinery to answer that question, my free AI Readiness Score takes about ten minutes — 20 questions across pilots, data, talent and governance, including the ones this essay turns on: whether AI work has a named owner, and whether anyone can say what the capacity is for.

Read the essay online at <https://burakgaliba.com/blog/the-ai-capacity-fork>

I built a free AI Readiness Score — 20 questions across pilots, data, talent and governance, about ten minutes, no email required: <https://burakgaliba.com/readiness>

© 2026 Burak Galiba · <https://burakgaliba.com> · Views are my own.