12

WHITE PAPER · PORTFOLIO & AI GOVERNANCE

Oversight Capacity, Not Model Capability, Is the Ceiling on AI Scale

Why the binding constraint on scaling AI is how much humans can supervise — and how to design for it

Marco Policani

Enterprise Portfolio · PMO · AI Operating Governance

policani.net · linkedin.com/in/marcpolicani

July 2026

DECISION SIGNAL

AI scale is capped by the human attention available to supervise its work.

Executive premise

There is a widely held assumption that the limit on how far an organization can scale AI is the capability of the models: get better models, deploy more of them, and output rises accordingly. That assumption is increasingly wrong, and getting more wrong every quarter. Model capability is abundant and cheapening. What is scarce, and what actually caps safe scale, is oversight capacity: the finite human attention and judgment available to review, correct, and answer for what the AI does.

This is not a contrarian framing; it is now the mainstream view among the people studying agentic operating models. McKinsey's analysis of the agentic organization states it directly: the scale of agentic adoption will be capped by how much oversight capacity humans can provide, making governance itself a potential bottleneck to productivity.[1] The same research offers a striking early-adopter ratio, a human team of two to five people supervising an agent factory of 50 to 100 specialized agents, which is impressive and, read carefully, is also a statement about a ceiling.[1] There is a number of agents beyond which a given human team can no longer answer for the work, and adding agents past that point does not scale output; it scales unmanaged risk.

The claim of this paper is that oversight capacity is the real constraint on AI scale, that it can be measured and designed for, and that treating it as a first-class resource, budgeted and engineered like any other, is what separates organizations that scale AI safely from those that scale their exposure. The rest of this paper defines oversight capacity, explains why it became the binding constraint, and lays out how to raise the ceiling without pretending it does not exist.

Why capability stopped being the constraint

For most of the recent history of enterprise AI, capability genuinely was the limit. Models could not reliably do the task, so the work of adoption was making them good enough. That era is ending for a widening set of tasks. As models handle more, more reliably, the bottleneck moves downstream, to the human who has to check the output, own the decision, and answer when it is wrong. Capability created the output; oversight has to absorb it, and oversight did not get cheaper at the same rate.

The asymmetry is the whole point. Generating output has become dramatically more scalable, because a model can produce more at near-zero marginal cost. Supervising output has not, because it still runs through a human whose attention is fixed and whose judgment does not parallelize. When one side of a system scales exponentially and the other does not, the non-scaling side becomes the constraint. In enterprise AI, the non-scaling side is human oversight, which is why capability, however impressive, is no longer where the limit lives.

This reframes what a serious AI scaling strategy is even about. The instinct is to plan for more and better models. The more accurate plan is to ask how much oversight the organization can bring to bear, how to make each unit of oversight cover more, and how to design work so that the oversight required per unit of output falls. An organization that adds models without answering those questions is planning to scale the thing that was already abundant while ignoring the thing that was already scarce.

What oversight capacity is

Oversight capacity can be defined precisely enough to manage:

Oversight capacity is the total human attention and judgment an organization can reliably apply to reviewing, correcting, and being accountable for AI output, across all the AI it is running, without the quality of that oversight degrading. It is finite, it is shared across every AI system competing for it, and it is consumed whether or not anyone is measuring it.

Two features make it the binding constraint. It is finite because human attention is finite: a person can genuinely own only so many decisions before the ownership becomes nominal. And it is shared: every AI system an organization deploys draws on the same pool of human judgment, so a new agent added in one function quietly reduces the oversight available everywhere else. Organizations rarely track this pool, which is exactly why they exceed it without noticing, until the oversight is so thin that it is oversight in name only, and the first evidence of the shortfall is an incident no one caught.

The span of supervision

The concrete planning quantity that follows from oversight capacity is the span of supervision: how many agents, decisions, or units of AI output a single accountable human can genuinely own. It is the AI-era analogue of span of control in management, and it deserves the same place in operating design. McKinsey's observed two-to-five people per 50-to-100 agents is a span of supervision, and while the specific ratio is an early-adopter figure rather than a law, the concept is durable: there is a number, and it is not infinite.[1]

The span of supervision varies with the work, and designing for it means being honest about what drives it. A high-stakes, high-ambiguity process where errors are costly and hard to detect supports a small span, a human can genuinely own only a few such agents. A low-stakes, well-bounded process where errors are cheap and obvious supports a large span. Treating the span as a single number across all AI is the error; treating it as a variable to be estimated per workflow, and then holding the number of live agents within it, is the discipline. The practical test is simple and uncomfortable: for each agent, can a named human actually answer for what it did this week? When the honest answer becomes no, the span has been exceeded.

Raising the ceiling with layered oversight

The span of supervision is not fixed, and the most important move in scaling AI is raising it without lowering the standard. The leading approach is to use technology to extend human oversight rather than replace it. McKinsey describes embedding control agents into workflows: critic agents that challenge outputs, guardrail agents that enforce policy, and compliance agents that monitor regulation, with actions logged and explainable.[1] Deloitte describes the same idea as agents guarding other agents that are then guarded by humans.[2] Layered this way, a small human team can answer for a much larger population of agents, because the routine supervision is itself automated and the humans concentrate on the exceptions and the accountability.

The public standards supply the operating actions that make layered oversight credible rather than notional. The NIST Generative AI Profile calls for continuous monitoring, red-teaming, independent evaluation, incident response, and structured feedback across the AI lifecycle, and attention to the human-AI configuration itself.[3] These are the mechanisms that let oversight scale: monitoring catches drift without a human watching every action, red-teaming finds failure modes before they reach production, and incident response bounds the damage when something slips. The NIST AI Risk Management Framework's govern, map, measure, and manage cycle provides the frame that keeps this an ongoing discipline rather than a launch checklist.[4]

The essential caveat is that layered oversight raises the ceiling; it does not remove it. Every layer of oversight agents ultimately answers to a human, and the critic and guardrail agents are themselves systems that can fail and that consume some oversight to maintain. Automated oversight is a force multiplier on human capacity, not a substitute for it, and an organization that forgets this simply relocates the ceiling to the supervision of its supervisory agents. The point is to push the ceiling as high as the work safely allows, while remembering it is still there.

Oversight is not reviewing everything

A common misreading of this argument is that it demands humans review every AI action, which would make AI pointless by capping its output at human throughput. That is not oversight; it is a bottleneck dressed as diligence. Effective oversight is not line-by-line review. McKinsey's framing is that humans move above the loop, defining policies, monitoring outliers, and adjusting the level of involvement rather than inspecting every action.[1] The skill is calibrating how much oversight a given workflow actually needs, enough to manage the risk, not so much that it pulls the AI back to human speed.

This calibration is where much of the real oversight-capacity gain lives. Uniform, heavy review of everything wastes the scarce resource on low-risk output that did not need it, leaving too little for the high-risk output that did. Risk-proportioned oversight, sampling low-stakes output, concentrating human judgment where errors are costly or hard to detect, and escalating the genuinely uncertain cases, stretches a fixed pool of human attention across far more AI. Getting this balance right, McKinsey notes, is what separates the organizations that capture the agentic advantage from those that either drown in review or lose control.[1]

Reading the constraint

Leaders can tell whether they are managing the real constraint by what they reach for when they want to scale AI. The table pairs the capability lever, which is largely exhausted, with the oversight lever that actually moves safe scale.

The capability lever	What it assumes	The oversight lever that actually scales
Deploy a better model	The limit is model quality	Raise the span of supervision per human
Add more agents	More agents means more output	Add oversight capacity, or the output is unmanaged
Trust the model more	Better models need less review	Proportion oversight to risk, not to trust
Review every output	Diligence means checking everything	Move above the loop; sample and escalate by risk
Buy more capability	Capability is the bottleneck	Layer oversight agents under human accountability

The right column is the same reframing five times: stop treating capability as the limit and start treating oversight capacity as the resource to be raised, proportioned, and budgeted. Every left-hand move fails the same way, by adding output the organization cannot answer for.

Designing for oversight capacity

Making oversight capacity a first-class design variable comes down to a short discipline that any AI program can adopt.

- Estimate the span of supervision per workflow, driven by stakes, ambiguity, and how detectable errors are — not a single number across all AI.
- Budget oversight as a real, finite resource, and charge every new AI deployment against the shared human-attention pool it will consume.
- Hold the number of live agents within the oversight available, so live AI tracks the humans who can answer for it, not the other way around.
- Proportion oversight to risk: sample low-stakes output, concentrate judgment where errors are costly or hidden, escalate the uncertain.
- Layer automated oversight — critic, guardrail, and compliance agents — under named human accountability to raise the span without lowering the standard.
- Instrument oversight with monitoring, red-teaming, and incident response, so degradation is caught by the system rather than by a human watching everything.

None of this slows a serious AI program in a way that matters. It front-loads the question that an over-scaled program pays for later, at a worse time, when there are more agents live than anyone can answer for and the first sign of the shortfall is a failure no one caught in time.

What it looks like in practice

The safest illustrations are operating patterns, not named systems. In governance work, the durable pattern is that AI scales cleanly only when oversight is designed alongside it, and stalls or fails when oversight is assumed to be free.

In one pattern, an organization scaling an AI-assisted process treated oversight as the planning constraint from the start: it estimated how many cases a reviewer could genuinely own, held the throughput within that span, and added reviewers or automated checks deliberately as it grew. The AI scaled because its oversight scaled with it, and the quality held because no human was ever responsible for more than they could actually see.

In a second pattern, a program scaled the AI first and the oversight later, on the assumption that better models needed less supervision. Output rose and so did unowned risk, until an error that no one had the capacity to catch surfaced downstream. The corrective was not a better model; it was rebuilding the oversight the scaling had outrun, proportioning it to risk, and holding future growth to the oversight available. The lesson repeats: capability was never the thing in short supply.

Evidence gaps this paper cannot close

Two limits should be explicit. First, the specific ratios in the literature, such as the two-to-five humans per 50-to-100 agents figure, are early-adopter observations from selected engagements, not validated benchmarks that transfer across industries, risk levels, or regulatory contexts.[1] They establish that a ceiling exists and is surprisingly reachable, not what the safe number is in any given setting; readers should measure their own span of supervision rather than import a headline ratio. Second, the return on oversight investment is not yet well quantified: the literature is clear that oversight caps safe scale, but it does not tell a leader precisely how much oversight a given AI portfolio needs to stay safe. Until that evidence matures, the defensible posture is conservative, budget oversight generously, measure the span you can actually sustain, and let live AI follow it.

Oversight capacity is a portfolio decision

Because oversight is finite and shared, deciding where to spend it is a portfolio decision, not a local one. Every AI system an organization runs competes for the same pool of human judgment, which means the question is never simply whether a given AI deployment is worth doing in isolation; it is whether it is worth the oversight it will consume, given everything else that oversight could cover. An organization that approves AI deployments one at a time, each justified on its own merits, will systematically over-commit its oversight, because no single approval ever sees the shared constraint it is drawing down.

This connects oversight capacity directly to funding discipline. Just as a portfolio has to hold its commitments within the capacity to deliver them, an AI portfolio has to hold its live systems within the oversight capacity to supervise them. The scarce resource has simply changed from delivery capacity to oversight capacity, and the discipline is the same: track the pool, charge every new commitment against it, and be willing to say not yet when the oversight is not there. An AI program that cannot say not yet on oversight grounds is not governing its scale; it is discovering it.

Framed this way, the highest-value oversight investments are often not more review but better allocation. Moving oversight from a low-risk system that was being over-watched to a high-risk system that was being under-watched raises safe scale without adding a single reviewer, exactly as reallocating capacity in a project portfolio raises throughput without new hiring. Leaders who feel they

are out of oversight capacity are frequently spending what they have on the wrong AI, and the first move is reallocation, not expansion.

The failure signature of exceeded oversight

Exceeding oversight capacity has a recognizable signature, and naming it lets leaders catch the problem before the incident that usually announces it. The first sign is nominal ownership: agents and AI systems that technically have an owner on paper, but whose owner could not actually reconstruct what the system did this week if asked. When ownership becomes a name in a document rather than a person exercising judgment, oversight has already thinned past the point of being real.

The second sign is review that has quietly become rubber-stamping. When the volume of AI output outruns the human capacity to examine it, review does not stop; it degrades into approval, with the human signing off on work they did not truly inspect because inspecting it all is no longer possible. This is more dangerous than no review, because it manufactures a false record of oversight that makes the system look governed while it is not. The third sign is surprise: incidents that emerge from parts of the AI estate no one was really watching, discovered downstream rather than caught upstream, which is the direct operational symptom of oversight spread too thin.

The corrective for all three is the same and is uncomfortable because it means slowing something down: reduce the live AI footprint to what current oversight can genuinely cover, or add oversight before adding more AI. An organization that responds to these signs by demanding better models is treating a symptom in the wrong system. The shortfall is in oversight, and only oversight, restored or reallocated, fixes it.

The counterintuitive economics

The oversight constraint produces an economics that runs against intuition and is worth making explicit, because it changes where the returns on AI investment actually come from. The intuitive view is that the value of AI rises with the capability and quantity of the models: more and better AI, more value. The oversight view corrects this. Beyond the point where oversight is the binding constraint, the value of additional model capability is capped by the oversight available to use it safely, so spending on more capability yields little while the oversight ceiling holds. The marginal dollar is better spent raising the ceiling than raising the capability under it.

This means the highest-return AI investments, past a certain maturity, are frequently investments in oversight rather than in models: in the monitoring, evaluation, and layered control that let a fixed human team safely answer for more AI, and in the workflow redesign that lowers the oversight each unit of output requires. These are unglamorous line items compared with a new model, which is precisely why

they are under-funded and why the organizations that do fund them pull ahead. They are buying the scarce resource instead of more of the abundant one.

It also means that the right measure of AI maturity is not how much AI an organization is running but how much it can run while still answering for all of it. An organization supervising a large AI estate with genuine, non-nominal oversight has achieved something far harder and more valuable than one running the same estate on oversight it has quietly outrun. The first has raised its ceiling; the second has merely postponed the moment it hits one. Measured honestly, safe scale, not raw scale, is the number that reflects real capability, and it is governed by oversight capacity, not by the models.

Oversight capacity is made of people

It is easy to discuss oversight capacity abstractly, as a pool of attention, and miss that the pool is made of specific, experienced people. Oversight is not a generic resource that can be summoned; it is the judgment of humans who understand the work well enough to know when the AI is wrong, which is a capability that takes years to build and cannot be conjured on demand. This makes oversight capacity not only finite but slow to grow, and it ties the ceiling on AI scale directly to the depth of experienced judgment an organization has.

The connection carries an uncomfortable warning. Organizations under pressure to show AI savings sometimes reduce exactly the experienced staff who constitute their oversight capacity, on the theory that AI now covers the work those people did. But the work those people did included the judgment that supervises the AI, and cutting them lowers the oversight ceiling at the very moment the AI is being scaled up against it. The result is an organization running more AI with less capacity to answer for it, which is the precise condition this paper warns against, arrived at through a cost decision that looked unrelated to oversight.

The design implication is that building oversight capacity is partly a workforce-development problem, not only a tooling one. The people who can supervise AI, who have the domain judgment to catch its errors and the context to own its decisions, are an asset that has to be grown and retained deliberately, especially the experienced staff and the developing pipeline behind them. An organization that treats these people as a cost to be minimized is depleting the reservoir that its AI scale depends on, and it will discover the shortage the way oversight shortages are always discovered: as an error no one had the capacity to catch.

This reframes a familiar tension. The question is not whether to invest in AI or in people; it is that scaling AI safely requires investing in the people who provide its oversight, because they are the ceiling. The organizations that scale AI furthest will be the ones that grew their oversight capacity, the experienced human judgment, in step with their AI, rather than trading one for the other and wondering later why the scale became unmanageable.

The shift

The organizations that scale AI furthest will not be the ones with the best models. On that dimension the market is converging, and capability is becoming a commodity input. They will be the ones that treated human oversight capacity as the real constraint, measured their span of supervision, raised it deliberately with layered oversight, proportioned it to risk, and refused to let live AI outrun the humans who answer for it.

Capability builds the output. Oversight decides how much of it the organization can safely own. The second is the scarce resource, and no quantity of the first substitutes for it. Treat oversight capacity as a first-class resource, engineer it as deliberately as the models, and the ceiling on AI scale rises as far as the work safely allows. Ignore it, and the ceiling arrives anyway, unannounced, as the moment the organization can no longer answer for what its AI just did.

Notes and sources

1. Alexander Sukharevsky, Alexis Krivkovich, Arne Gast, Arsen Storozhev, Dana Maor, Deepak Mahadevan, Lari Hamalainen, and Sandra Durth, "The agentic organization: Contours of the next paradigm for the AI era," McKinsey & Company, September 26, 2025. Verified July 10, 2026. https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/the-agentic-organization-contours-of-the-next-paradigm-for-the-ai-era

2. David Mallon, Brad Kreit, and Natasha Buckley, "Rethinking operating models for humans with agents," Deloitte Insights, April 2, 2026. Verified July 10, 2026. https://www.deloitte.com/us/en/insights/topics/talent/operating-models-for-humans-ai-agents.html

3. National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," NIST AI 600-1, July 2024. Verified July 9, 2026. https://doi.org/10.6028/NIST.AI.600-1

4. National Institute of Standards and Technology, "AI Risk Management Framework Core," excerpt from AI RMF 1.0, 2023. Verified July 5, 2026. https://airc.nist.gov/airmf-resources/airmf/5-sec-core/

About the author

Marco Policani is an enterprise portfolio, PMO, and AI operating-governance leader. He builds portfolio governance systems where AI carries the analytical load and named humans own every decision — an approach documented across case studies, walkthroughs, and working governance guides on his portfolio site.

Portfolio & governance library: policani.net/governance · Full portfolio: policani.net · LinkedIn: linkedin.com/in/marcpolicani