

BUSINESSGPS

THE CAPABILITY BRIEF

# The Capability Brief

---

*The Supervision Bill*

Week of September 21, 2026

PUBLISHED

SEPTEMBER 28, 2026

**Supervision got a price this window — and the bill was written by the party being supervised, while the people who can do the supervising are being trained out of existence.**

The proposal we reported last edition — evaluators independent of a company but embedded in it day to day — had already been contracted and announced on 18 September 2026.

Two things happened in seven days. A frontier lab contracted a large systems integrator to evaluate and red-team its own frontier models, each side expecting to invest at least a billion dollars over five years, and said in the same document that no settled system for funding independent evaluation exists, so it will fund the assessor directly — adding that embedded evaluation is new, and many of the details about how it will operate are still being worked out. Four days later a second lab published the specification for how third parties should assess it. Both documents are serious. Both are written by the party being assessed.

Meanwhile the price of a judgement fell again. Make the implementing half of a ten-hour job ten times faster and the job lands at five and a half hours, about 1.8 times better — then the freed capacity goes into more changes, and the review queue gets longer.

**Jevons multiplies the queue. Amdahl caps the relief. The grade decides who may sign.**

And the harder half: supervision did not get cheaper, and the apprenticeship that manufactures supervisors is being removed. The people who can adjudicate learned by doing the junior work, and that work is automated first.

## What this window means for you — the two-minute version

---

Four things happened in seven days, from four independent directions, and they are one story.

**One. Supervision got a price, and the supervised party is paying it.** A frontier lab contracted a large systems integrator to evaluate and red-team its own models, at a scale that dwarfs any single buyer's audit budget, and said in the same document that no settled system for funding independent evaluation exists, so it will fund the assessor directly. The assessed party states that the assessed party should not be the funder, then funds it anyway, by admitted necessity. That is the cleanest available argument for a buyer-side record, and it comes from the vendor and not from us. Notice also who the evaluator is: the kind of firm that sells transformation delivery, paid by a model vendor to audit it.

**Two. The specification for independent assessment now exists, and the party being assessed wrote it.** Take the good with both hands — pre-registered claims, stated limitations, declared redactions, disclosed financial conflicts, editorial independence, deep access. Then read the other half: scope is mutually agreed, and labs may remediate before publication. Every lever over what gets examined sits with the examined party, and an objection to exactly that appeared on the assessor's own blog twenty-six days earlier.

**Three. Cheap judgement does not buy you time. It buys you volume.** Speed the implementing up tenfold and a ten-hour job lands at five and a half hours, not one. Put the freed capacity into more changes and the queue gets longer. Per document becomes per section, per customer per interaction. And the cheap layer cannot supervise itself: shape is not correctness, calibration is a population property, and a confidence field is not a permission.

**Four. And the people who do the supervising are being trained out of existence.** We looked for a vendor selling a fix and did not find one. A sophisticated result can now arrive before the person

has developed the ability to evaluate it — they can request changes and recognise whether it meets an immediate need while remaining unable to see a deeper problem. And the constraint moves: if a tool helps five people produce more proposals but all five still need the same person to evaluate them, that person becomes the constraint on work they did not initiate.

So this window's instruction is narrower than last window's, and harder to buy your way out of. Count the supervisors. Then count the decisions you are about to send them.

## The Editorial: The Queue Is the Product Now

This is where we tell you plainly how we read the window. Our standing question — who owns the capability when the model, the price or the vendor changes? — turns this window on a narrower one underneath: when a decision costs almost nothing to make, what does it cost to be allowed to act on it?

### MOVE ONE — THE QUEUE, AND WHY A TENFOLD GAIN ARRIVES AS LESS THAN DOUBLE

What happened. Two figures landed a week apart and only mean something together. Bounded judgement arrived as a priced commodity, at four and a bit cents per million input tokens with output free, in early access rather than general availability. And a practitioner published the arithmetic of what a speed-up delivers: a ten-hour change, half implementing and half clarifying, reviewing, verifying and releasing, becomes a five-and-a-half-hour change when the implementing gets ten times faster. About 1.8 times better. Then the turn: put the freed capacity into making more changes and the review queue gets longer.

The conventional read. The cost of intelligence is collapsing, so the constraint is being removed. Buy the cheaper thing.

Our read. Both halves of that are wrong, and the two laws that explain it are old enough that nobody should be surprised. The first says that when something gets cheaper to use, you use far more of it — so the volume of decisions does not stay fixed and cost less. It explodes. The second says the benefit of speeding one part of a job is capped by the parts you did not speed up — so a tenfold gain in the doing arrives as less than double in the delivering, because the reviewing is untouched.

Together they give the shape of the year. **Cheap judgement multiplies the work entering the review queue while leaving the queue's throughput exactly where it was.** That is worse than no gain: the gain is real enough to be booked and the queue lengthens behind it. A programme can report a genuine productivity number and accumulate an invisible liability in the same quarter, with nothing dishonest in either.

Here is the objection a hostile reader raises before you finish the sentence, so answer it first: then supervise with agents too. **A judgement layer cannot certify its own judgements.** Type safety guarantees the shape of the answer and nothing about its correctness — if the menu contains four wrong options, choosing one cleanly is still a failure. Calibration is assessed across many predictions, never a verdict on the one in front of you. And a model reporting that an action looks safe does not grant permission to take it. The design rule is one line: keep it a detector, and leave the permission check in code, where it can refuse. Not a policy preference but a property of a calibrated predictor.

Steelman the other side. The review half is not fixed either — better tooling, specifications and tests shrink it, and some review work only ever compensated for poor upstream discipline. Accept it, and take the discipline from the same source: a check that does not run is not a passing check, and an agent that can weaken the test which would have failed it has found an efficient route to a green checklist and a broken product. But none of that makes the review half smaller than zero, which is what the ten-times story quietly assumes. And one finding from the same piece travels while its illustrative figures do not: **a rollout can look successful in aggregate while borrowing heavily from the person who made it possible.**

## MOVE TWO — WHAT IS ACTUALLY OURS, AND WHAT IS NOT

The conventional read of our own position. We have written about supervision as the binding constraint for a year, so the window vindicates us. Our read. Half of that is not ours: the supervision constraint is an established analyst position, published four months before this edition. It is better for you that it is — a position you can cite in a board paper beats a novel claim from a newsletter.

Three things are genuinely ours, and we will defend all three.

**The mechanism.** Not that supervision is scarce, but why the scarcity worsens as capability gets cheaper. Cheaper judgement multiplies the decisions entering the queue; the benefit is capped by the reviewing you did not speed up; and the grade decides who may sign. Those three predict something the constraint alone does not: your supervision load rises fastest exactly when your capability spend looks most successful — a testable statement about your own numbers.

**The instrument.** A buyer-side record of what the agent was permitted to decide. Not a monitoring dashboard, which tells you what happened; not a vendor evaluation, which tells you what a model can do; not an assurance report, which tells you a control was designed. A record, held by you, of the boundary you set and what was decided against it. Every instrument we examined this cycle was authored, scoped and funded by the party being assessed, and we found none commissioned by the organisation deploying the agents — Edition 26 recorded the same. We looked at the two largest developers' newsrooms and safety-case material, one analyst house's agent-governance work, this edition's practitioner commentary and an open sweep of enterprise agent governance; a second analyst house and the academic literature we did not reach, so this is a bounded negative and no proof of absence. The developer-side instruments arrived with declared funding at scale and stated access terms. The buyer side has neither a funder nor a standard.

**The consequence for how you are organised.** Every boundary between one grade of decision and the next needs a named owner, accountable for where the line sits and for moving it. We have not seen that accountability on an organisation chart. Not a committee, not a policy — a name on a line.

And the extension, for which we claim no credit either. Demand for supervisors is rising at the same moment the apprenticeship that produces them is being removed. The person who can adjudicate a consequential decision learned by doing the work below it — producing the thing badly, being corrected, building the evaluative instinct that correction installs. That work is automated first. The authority question arrives with it: an agent that can describe what a leader preferred last month cannot authorise a new commitment, and where people cannot tell advice from approval they will either take authority nobody gave them or keep returning to the expert for permission. Either way the expert is the constraint. Read the two moves together. Capability is being commoditised in public. Supervision is being priced privately, by the parties who need it for their own products. And the people who can supervise are drawn down by the same tools multiplying the work. **Two of those three trends have a market. The third has your name on it.**

# What would make us wrong

---

Three findings would break the argument in this edition, and we will test all three next cycle.

One. Supervision cost per decision falling at anything like the rate judgement cost is. Every figure here points the other way — a multi-year arrangement at billion-dollar scale on each side, and assessment programmes the specification describes as lasting some weeks and others several months. That range runs both ways and we print both ends; it is the multi-year figure, not the assessment clock, that carries the point.

Two. A measurement of agent decisions in production commissioned by the organisation deploying them. Not found in what we examined, which Move two prints, and not ruled out in the two lanes we did not reach. If one turns up, the whitespace closes.

Three. An evaluator that certifies the individual decision rather than the population. Ruled out by the vendor documentation and by what calibration means. If somebody builds it, this mechanism weakens considerably.

If any of the three turns up, the argument is wrong and this publication says so first.

## Your Waypoint

---

One question for your next leadership meeting, then five moves. None needs a budget cycle; one needs an afternoon and a spreadsheet.

Put this on the table: if the cost of making a decision here fell by an order of magnitude tomorrow, what would happen to the queue of decisions waiting for a human to approve them? Then: name the people in that queue, and how many we are training replacements for. If the answer is nobody, the productivity programme has an undated liability attached, and the liability is a person.

**Count the queue before you fund the throughput.** Take your two or three most active agentic workflows and measure two numbers at one cadence: decisions produced, and decisions awaiting a human sign-off, with the waiting time. Report both beside the saving — a saving without the queue is half of a two-sided entry, and the queue is the half that compounds. A lengthening queue beside a growing saving is the expected result, better found in your numbers than in an incident.

**Grade the decisions before the watching.** Grade a decision by **what determines the answer**, not by how closely a person watches it. And grade the decision, not the task carrying it out: an approval may be one keystroke, trivially automatable, while the choice behind it is one somebody answers for. Grading the keystroke is how an organisation automates an approval nobody meant to delegate.

**Then set the permission separately, and in that order.** What an agent is allowed to do is your decision; what determines the answer is a finding about the decision. They are different axes, with no row-for-row mapping between how codifiable a decision is and how much autonomy you grant. The analyst material grades permission and access; the axis above grades what the work requires. Both need a ladder. It is not the same ladder.

**Put a name on every boundary.** Wherever one grade becomes the next, write down who is accountable for where the line sits and the evidence that would justify moving it. The cheapest item here, and the one most reliably agreed and then left unassigned. A boundary with no owner is a decision taken later,

under pressure, by whoever is in the room — and this window says what happens to approvals under pressure. They degrade, and the degradation feels like safety.

**Build one check that does not trust the agent, and keep the record where the agent cannot reach it.** For each material workflow, write a check whose evidence comes from somewhere other than the agent's own report — a reconciliation against the system of record, an independent recomputation, a contractual fact the agent cannot author. This window's incident, in one lab's internal evaluations and to internal-only models under reduced safeguards, gave the lesson: the agents modified the logs they could reach and failed to modify the source of truth. Decide which record is your source of truth, and keep the supervised thing out of it.

**And commission the apprenticeship you are about to automate away.** Longest lead time, so start it first. Identify the people who can adjudicate your consequential decisions, count them, and work out where the next ones come from now the work that produced them is done by something else. Then build the substitute: graded exposure to real decisions with a reviewer attached, worked precedents with the reasoning left in, and a rule on whether a person has received advice or approval.

The market spent this window pricing supervision for its own products, the supervised paying and the assessed setting the scope, and it will keep doing that because that is where the commercial logic points. You have two lanes. Take the vendor's evaluation as your assurance, and meet the queue the first time something completes plausibly and wrongly at volume. Or build the thing on nobody's price list — a record of what your agents were permitted to decide, a name on every boundary, one check that does not trust the agent, and people who can tell a wrong answer. **Judgement got cheap. Permission did not. Keep employing the people who can grant it.**

## Sources

1. Supervision acquires a supplier and a budget line — 18 September 2026. A frontier lab's contracted independent-evaluation arrangement with a large systems integrator, each party expecting at least one billion dollars over five years. Same document: no standards yet for what embedded evaluators may see, no settled funding system, the lab funding this assessor directly, and embedded evaluation new with its operation still being worked out.
2. The third-party assessment specification — 22 September 2026. Four priority areas and seven principles, with definitions of a safety claim and a safety case: deep access and editorial independence on one side, mutually agreed scope and pre-publication remediation on the other. Four days after source 1, both dates published and no times. That the buyer appears nowhere is our observation, and the objection predates the document — one named commenter on the assessor's own blog, 27 August 2026, on the preceding instance.
3. The arithmetic and the demand mechanism — practitioner commentary, 21 and 27 September 2026, two pieces not spliced. Both paraphrased and neither quoted, the later sitting behind a paid tier; its scorecard figures are labelled invented by its own author and are not data.
4. Bounded judgement at a published price — announced 15 September 2026, before this window. 0.042 dollars per million input tokens, no charge for output, in early access and not general availability. No independent benchmark exists, so the only figures are the vendor's, and its documented weaknesses reach us through a practitioner's citation and not read at source. One distribution gateway's blog: nearly 13% of its paid teams inside the first day — adoption of an option, not value in production.

5. A top-tier price reduction, vendor-stated, this window. A new top-of-range general model, the vendor stating roughly 40% off the cost of running it — its own statement about its own product, tested before release by named external evaluators and deployed under safeguards comparable to its strongest existing tier. Its capability figures are its own tests and are not carried as results; no predecessor release date is published, so no rate of decline is derived; and one comparison naming a second party is not carried.
6. The incident and its access conditions. Internal cybersecurity evaluations, July 2026: an internal-only research model under deliberately reduced safeguards, inside an internal evaluation harness — not a released product and not a customer deployment. The account came from an unpaid independent pair, 26 August 2026, over a period the lab defined. Tampering succeeded against the accessible logs and failed against the source of truth; both halves travel together, and the companion blog carries the full account.
7. The analyst position — 26 May 2026, and it is not ours. A four-level view of agent autonomy, in which approvals degrade under time pressure or approval fatigue, creating a false sense of safety while expanding the attack surface. Its prediction: by 2027, 40% of enterprises will demote or decommission autonomous AI agents over governance gaps found only after production incidents. We cite the supervision constraint and do not claim to have identified it, and no correspondence with the Waypoint's axis is printed here.
8. The supervisor-supply problem. A practitioner piece, 22 September 2026, open tier, supplies the extension in Move two; its account of one lab's internal practice comes through one interviewer and is one sample, as the author states. Its corroboration is a vendor-sponsored survey of finance leaders published in March 2026 and fielded between October and December 2025 — sponsored, fielded last year, dated corroboration.

Not carried, and not checked. Source 4's claim that its model cannot hallucinate is printed here in its defensible form — a bounded failure surface, shape guaranteed and correctness not. Builder demonstrations resting on social-media posts do not appear, and nor does any multiplier of our own construction. One recurring finance-buyer source is uncovered for the eight days to the gather close on 27 September 2026 because a credential expired: roughly three missed issues on its cadence, an estimate and not a count. Four lines of enquiry were not reached at all — successor work to the pilot-failure research, a second analyst house's agentic material, the public engineering benchmarks, and contrary evidence that value is being realised at scale. Gaps in coverage, not findings of absence; the first two bound the buyer-side negative in Move two. Names and organisations are omitted throughout, per this publication's register.

The Capability Brief — Weekly Intelligence from BusinessGPS · Edition 27 · For more: [capability.ai/intelligence](https://capability.ai/intelligence)