Briefing

Test before you limit

An evidence-based path for the AI governance decision

lateralworks · FTTM advisory

A recommendation for any organization weighing a restriction on AI: run a bounded, measured experiment before setting policy that could slow products to market.

Prepared by
lateralworks
FTTM advisory

Date
June 2026
Briefing paper

Online
lateralworks.com
AI governance series

Table of contents

Test before you limit

Core thesis. The decision in front of any organization is whether policy gets written from evidence or from fear. A bounded experiment with a hard go/no-go gate lets a council govern AI responsibly and learn what it is worth, before a restrictive default quietly slows delivery and pushes AI use into channels no one can see.

Overview

Executive summary

An organization weighing this question is about to make a policy decision, not a technology decision. The question is not whether artificial intelligence is safe in the abstract. It is whether to halt AI inside the working methodology until policy catches up, or to let a small, measured experiment show what it is worth and write the policy from the result.

The stakes are concrete. An organization on an aggressive multi-product schedule has the least room to give up leverage. The lateralworks methodology uses AI to compress planning and tracking. It synthesizes many data streams into the plans and decision models the core team works from. Removing it does not return the organization to neutral. It strips out a force multiplier at the moment the schedule needs one, and it adds the headcount and timeline pressure leadership has already flagged.

The evidence on restrictive AI policy is now substantial, and it points one way. Organizations with stronger AI governance adopt faster, not slower [4]. Field experiments confirm the productivity gain is real and large, and that it depends on a human staying in the loop [10, 11, 12]. Restriction does not stop AI use. It pushes it into unmanaged “shadow” channels [5, 6]. And the fear-driven pause has a measured cost: most organizations have already delayed deployments on security grounds, and those delays create risks of their own [5].

The council’s real concern, how the organization’s data is handled, is legitimate and answerable. lateralworks runs on a configuration where conversations are not used to train models, and stronger contractual and isolated postures are available when the data classification calls for them [8]. That is a question to engineer, not a reason to stop.

The recommendation: do not impose a full stop. Run an eight-week pilot on two live programs, governed by a simple three-lane framework, measured against criteria agreed up front, and ended by a hard go/no-go gate. Let evidence set the policy, rather than fear or hype. This briefing gives the council the framework to do exactly that.

01

The decision

A policy choice, not a tech choice

Leadership can frame this decision precisely for a council. Read strictly, a restrictive policy conflicts with how the lateralworks tools work, and absent a conversation the default is a full stop on all AI touching company data. That is the fork: a full stop now, or a controlled path that earns its policy.

This is not a referendum on whether AI is safe in general. It is a choice about how an organization learns what AI is worth in its own product-development process: by testing it under controls, or by ruling on it before the evidence exists.

What is actually being governed

The lateralworks tools use AI for synthesis: pulling schedules, status updates, decision inputs, and economic parameters into one organized structure that feeds the plans and models the team scrubs together. In practice that spans planning and tracking, decision models, cost-of-delay analysis, schedule evaluation, and weekly status reports. The companion paper, *How lateralworks uses AI*, documents each of these in detail. The point for the council is narrow. This is assisted synthesis with a human deciding at the end. It is not autonomous decision-making, and the data is not sold or surrendered.

Why the timing makes this urgent

An aggressive multi-product schedule is the reason a force multiplier matters now rather than later. Withdrawing AI from the methodology at the start of that push removes leverage precisely when the program can least absorb the loss. Leadership has already named the consequence: a significant additional headcount ask and likely timeline impact across most of product development if the restrictive policy holds. The cost of a wrong “no” here is measured in schedule, not in abstractions.

The question the council asked

The right question for a council to ask is this: what is the decision criteria for allowing or rejecting something? That question, not a blanket ban, is the correct basis for policy. Section 04 answers it with a framework the council can adopt directly.

02

The evidence

Governance and speed move together

The instinct under uncertainty is to slow down. The data of the last two years says that instinct, applied as a blanket restriction, produces the opposite of safety. It yields slower delivery *and* more unmanaged risk.

None of what follows is a lateralworks opinion. It is the consistent finding of independent enterprise surveys and peer-reviewed field experiments across 2023–2026.

AI use is already the norm

Seventy-eight percent of organizations reported using AI in 2024, up from 55 percent a year earlier [1]. Boards are catching up: the share reporting AI was not on their agenda fell from 45 percent to 31 percent in a single year [2]. Yet only about a third of companies have a formal AI policy at all [3]. The gap between how widely AI is used and how little of it is governed is exactly where risk lives. A blanket ban widens that gap rather than closing it.

Stronger governance correlates with faster adoption

The clearest finding cuts against the fear: organizations that build stronger AI governance adopt AI *faster* [4]. The confidence gap is stark. Only 7 percent of pilot-stage organizations are confident they could pass an independent AI governance audit within 90 days, against 74 percent of those that have fully integrated AI [4]. The same research is blunt about the failure mode: centralized review bodies that try to clear every use case get overwhelmed and become bottlenecks that slow the business without reducing risk.

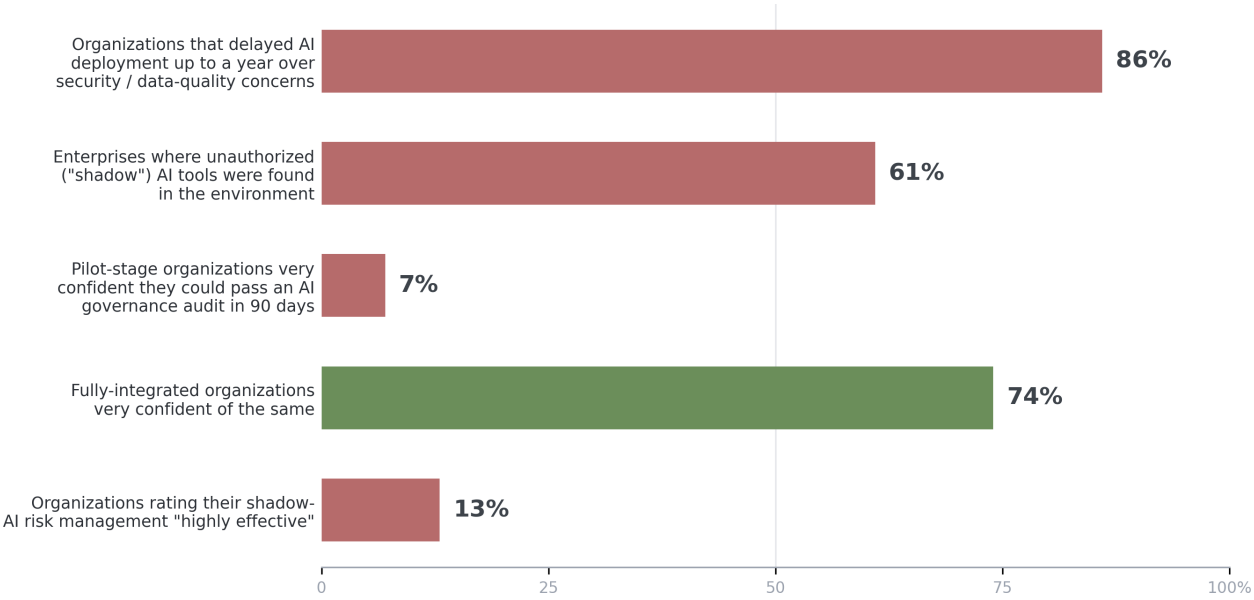


Figure 1. The cost of a restrictive default. Independent surveys, 2024–2026, show delay and shadow use rising where governance lags or over-restricts, and confidence rising where AI is properly integrated [4, 5, 6].

Restriction does not stop AI, it hides it

Eighty-six percent of organizations have delayed AI deployments by up to a year over security and data-quality concerns [5]. The demand does not disappear when the sanctioned path closes. People bring their own tools instead. Sixty-one percent of enterprises have found unauthorized AI tools in their environment, only 26 percent monitor AI use at all, and just 13 percent rate their management of shadow-AI risk as highly effective [6]. The mechanism is well documented: when governance is too slow or too restrictive, it forces shadow AI in the name of productivity [7]. A blanket “no” manufactures the exact unmanaged-data risk it was meant to prevent.

The productivity gain is real, and it is measured

The enterprise surveys describe the governance failure. A separate body of peer-reviewed field experiments measures the productivity itself, and the results converge. A Harvard study run with Boston Consulting Group across 758 consultants found that, on tasks inside the AI's capability frontier, consultants using AI finished about 25 percent faster and produced work rated roughly 40 percent higher in quality than the control group. On tasks outside that frontier they were more likely to be wrong, which is the reason human review is not optional [10]. A Stanford and MIT study of 5,179 customer-support agents found a 14 percent average productivity gain, rising to 34 percent for the least-experienced agents, as the tool spread the practices of the best performers to everyone else [11]. An MIT experiment with 453 professional writers, published in *Science*, found a 40 percent reduction in time and an 18 percent quality gain, with the time shifting out of drafting and into editing and judgment [12].

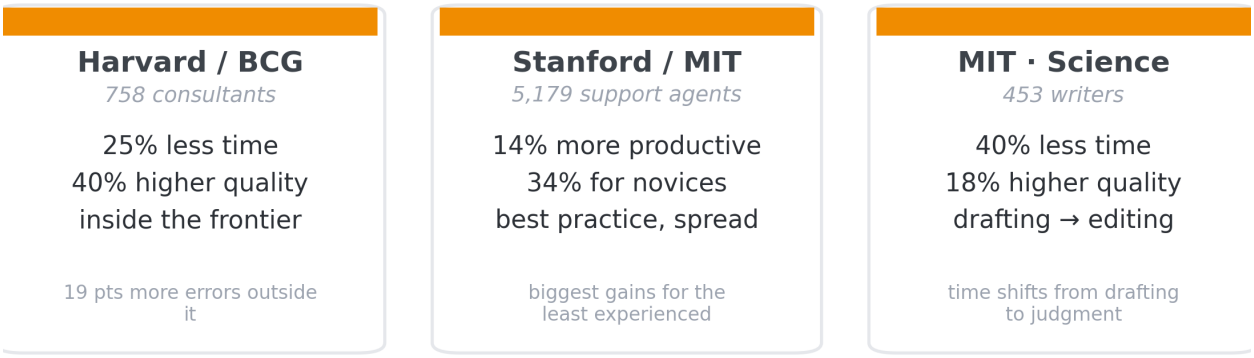
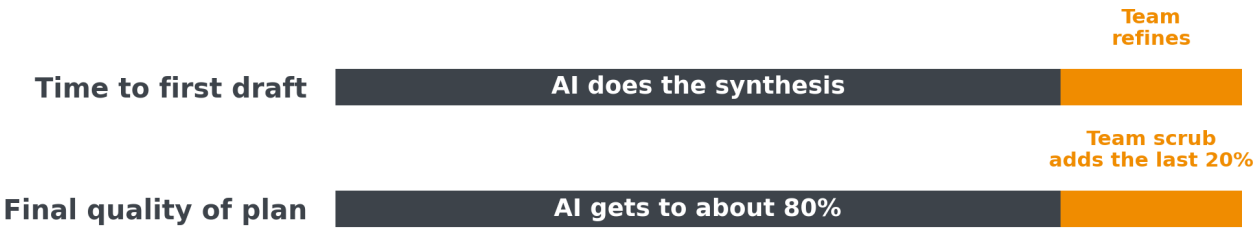


Figure 2. Three independent field experiments, 2023. Each measured a large productivity gain inside the AI's frontier. The gain is uneven, and it depends on the human keeping a clear role in review and direction [10, 11, 12].

AI to 80 percent, the team to the last 20

These findings describe exactly how the lateralworks flow is built. In the documented case study, AI did roughly 80 percent of the synthesis work, getting the plan to a usable draft in hours, and the team carried the last 20 percent: the judgment, calibration, and local context the model cannot supply [15]. The split is the operational form of the Pareto principle. Most of the value comes from a small share of the work, and AI makes that share nearly free, so the team can spend its full attention on what determines whether the plan is correct [13].

The discipline that turns the draft into a plan is iteration. The work moves through six stations. The first three, gather, distill, and synthesize, are where AI does the heavy lifting in minutes. The last three, organize, scrub, and refine, are human, and they cannot be shortened. The team reads the draft for factual errors, scope drift, and missing context, then feeds its corrections back as the next prompt. One pass is not enough. Three is usually enough [14, 15]. This is the same human-in-the-loop control the council is being asked to protect, already operating inside the methodology.



AI first cuts the time to a draft. The team adds the judgment AI cannot supply.

Figure 3. The 80/20 split. AI compresses the time to a first draft. The team scrub determines the final quality. The order matters: AI first, the team second [13, 15].

The lesson

Two findings sit on top of each other. Restriction does not make AI use safer; it pushes it into channels no one governs. And the productivity gain is real, large, and contingent on a human staying in the loop. The failure mode is not moving too fast. It is writing policy before the evidence exists, then finding the policy unenforceable because the work has already routed around it. The durable move is to make the sanctioned path fast and safe enough that no one needs a shadow one, and to decide its limits from what a controlled test actually shows.

The real risk

What a blanket no costs

A restriction does not return you to zero. It removes a force multiplier, and hides the use it cannot stop.

The pattern across enterprise AI, 2024–2026

Grant Thornton · AvePoint · Cato Networks

03

The real concern

Security is a question to engineer

A council’s real question is direct: how are the tools configured if company data goes into them, and what are the risks and benefits? That is the right question, and it has concrete answers. Treat it as an engineering decision about assurance level, rather than a binary about whether AI is “allowed.”

How lateralworks is configured today

AI work runs on Anthropic’s Claude under a configuration where the model-training setting is turned off. With that setting off, conversations are not used to train any model, and data is held under the standard short retention window rather than the extended one [8]. Anthropic publishes independent third-party security attestations, including SOC 2 Type II and ISO certifications, through its Trust Center, which a security team can request and verify directly [9]. The companion paper sets out the full data-handling posture.

Match the control to the data, not the fear

Not all data needs the same protection. The strongest answer to a sensitive-data worry is a posture that removes the training question by contract rather than by a consumer setting. Commercial and enterprise terms prohibit training by default. The API adds short or zero retention under a data-processing agreement. A firewall-isolated deployment keeps data on-premises entirely. lateralworks can operate at whichever of these rungs the organization’s classification requires, and can escalate the sensitive classes without slowing the rest.

What this means for the most sensitive data

Opening the most sensitive data to controlled, non-training environments under a scoped waiver is how mature programs handle this. Classify the data, then route each class to a matching control. The pilot can begin immediately on lower-sensitivity classes while the council sets the rule for the most sensitive material. Nothing about the schedule requires exposing crown-jewel IP to make progress this quarter.

04

Decision criteria

Three lanes, keyed to data sensitivity

The question of what the criteria should be for allowing or rejecting a use has a clean answer. Here is a framework a council can adopt directly. It replaces a single yes/no with three lanes, sorted by the sensitivity of the data involved and whether a human reviews the output before it acts. Most work falls in the first two lanes and should move without friction; only the genuinely sensitive cases need escalation.

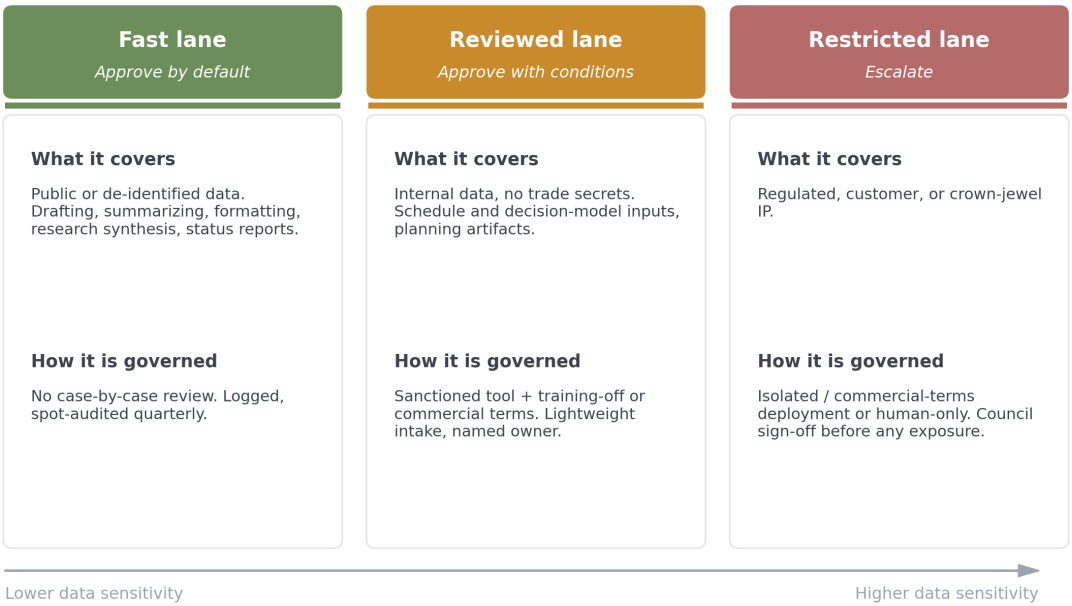


Figure 4. A three-lane decision rule. Each use is sorted by data sensitivity and reversibility into a fast, reviewed, or restricted lane, with a matching control. The rule is enforceable. A blanket ban is not.

How to apply it

Two questions decide the lane. First, how sensitive is the data? Second, does a person review the output before it is used? Public or de-identified data with a human in the loop is fast-lane: approve by default and audit on a sample. Internal data that holds no trade secrets, used through a sanctioned, training-off tool with a named owner, is the reviewed lane. Regulated, customer, or crown-jewel IP is restricted: council sign-off and an isolated or commercial-terms path before any exposure.

Why a tiered rule beats a blanket one

A single gate over every use becomes the bottleneck the evidence warns about [4, 7]. A tiered rule keeps the large volume of low-risk work moving and concentrates scrutiny where it actually belongs. It is also enforceable: people will follow a fast sanctioned path, whereas a blanket ban simply moves the work somewhere the council cannot see.

What to measure on the lanes

Track two numbers: the approval cycle time for a use, and the share of AI activity that sits inside the framework. If cycle time climbs or coverage drops, the policy is drifting toward bottleneck and shadow. Adjust it before it bites. A governance rule that cannot report its own cycle time is already the bottleneck.

05

Scoring the work

Weighing value against risk

The three lanes decide how data is handled. They do not, on their own, tell the council which applications are worth pursuing or in what order. That second question, weighing the value of an application against its risk, is where committees usually drift to gut calls or to whoever argues loudest. A simple, defensible scoring model fixes that.

Why a scoring model, not a gut call

The Analytic Hierarchy Process (AHP) is a standard decision method for exactly this problem: many competing criteria, no single obvious answer, and a group that needs to agree. It works in two steps. First, the council compares the criteria against each other in pairs, and the method derives a consistent set of weights from those comparisons. Second, each AI application is scored against the weighted criteria to produce one number the committee can rank and defend. The comparisons are explicit, so the logic is auditable rather than personal.

Six criteria, balanced across value and risk

lateralworks recommends six criteria, three that measure value and three that measure risk-control, each framed so that a higher score is always better for the company:

- **Time-to-market impact.** How directly the application compresses the critical path. For a program on an aggressive multi-product schedule, this is the prize.
- **Output quality uplift.** How much it improves the quality of plans, models, and decisions over manual work.
- **Effort leverage.** Hours returned to the team, multiplied by how often the application is reused across programs.
- **Data safety.** Higher when the data is less sensitive or the handling posture is stronger. This is the direct tie back to the three lanes.
- **Controllability.** Higher when a human reviews the output, the action is reversible, and the tool advises rather than acts.
- **Feasibility.** Higher when the application is cheap and quick to stand up and integrate into the workflow.

Deriving the weights by pairwise comparison

The council compares each pair of criteria on a 1-to-9 scale: equal, moderately more important, strongly more important, and so on. The matrix below is a worked starting point that leads with time-to-market while holding data safety and controllability close behind. The council should adjust the judgments to its own risk appetite, and the method then recomputes the weights.

	TTM	Qual	Effort	DataSafe	Control	Feasib
TTM	1	2	3	2	2	4
Qual	1/2	1	2	1	1	3
Effort	1/3	1/2	1	1/2	1/2	2
DataSafe	1/2	1	2	1	1	3
Control	1/2	1	2	1	1	3
Feasib	1/4	1/3	1/2	1/3	1/3	1

Table 1. Pairwise comparison of the six criteria (Saaty 1-9 scale). Read each row against each column: how much more important is the row criterion than the column criterion?

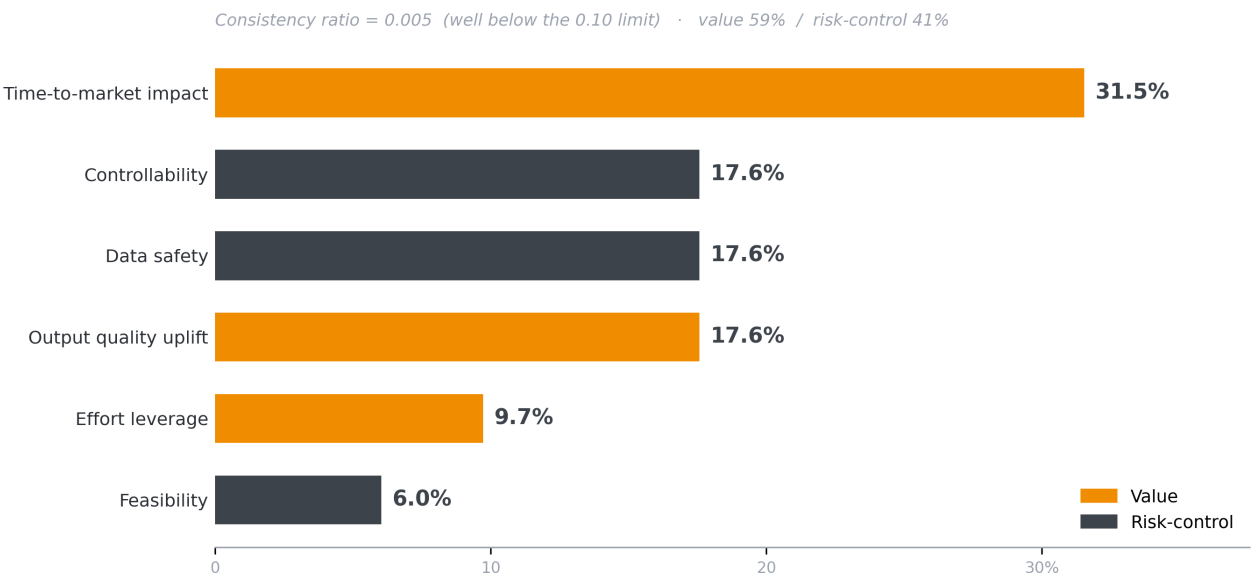


Figure 5. Criteria weights derived from the comparison matrix by the eigenvector method. Time-to-market leads at roughly a third of the weight, and value and risk-control split about 59 / 41. The consistency ratio of 0.005 is far below the 0.10 threshold, so the judgments hang together.

Scoring an application

Each application is then rated 1-to-9 on every criterion, and the ratings are combined with the weights into one global score. The ranking below applies the model to five AI applications already in or near a product-development workflow, plus two deliberate anti-patterns. The five legitimate uses cluster together. The two anti-patterns fall well below them. They do not lack value; the model docks them where it should. An autonomous tool that acts without review collapses on controllability, and sensitive customer data on an ungoverned tool collapses on data safety.

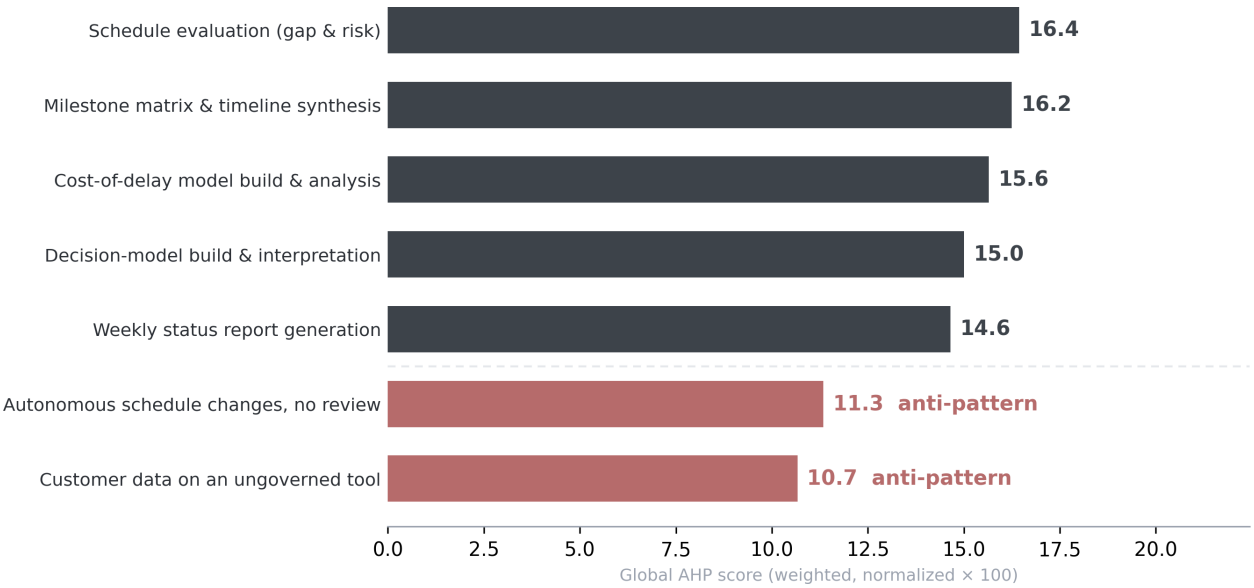


Figure 6. Worked example. Global AHP score for seven applications. The dashed line separates the legitimate cluster from the anti-patterns, which the weighting drives down precisely on the risk dimensions the council cares about.

How the council uses the score

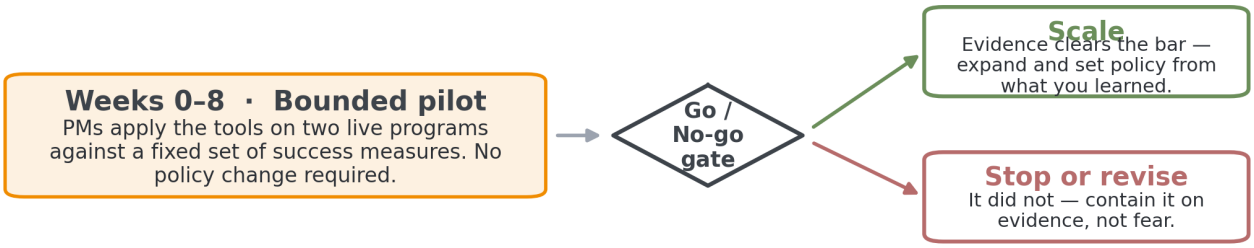
The score and the three lanes work together. A high score with low-sensitivity data lands in the fast lane and is approved by default. A high score on sensitive data does not get rejected. It routes to the reviewed or restricted lane, with the score guiding what conditions to attach. A low score is deferred regardless of how loudly it is championed. As with everything in this methodology, the model informs the decision; the council makes it. The number ends the argument about priority, not the human judgment about whether to proceed.

06

The recommendation

Run the experiment, then write the policy

lateralworks' recommendation to a council is specific: do not full-stop. Authorize a bounded, eight-week pilot on two live programs, governed by the three-lane framework, measured against criteria agreed up front, and ended by a hard go/no-go gate. Then write the policy from what the pilot shows.



The gate is decided on measures agreed up front, for example:

- Planning-cycle time vs. baseline
- Status-report hours recovered
- Schedule / decision-model rework rate
- PM-reported quality and trust
- Any data-handling incident

Figure 7. A time-boxed experiment with a hard gate. Eight weeks of measured use on two live programs, then a single evidence-based decision: scale and codify, or stop and contain, on data rather than fear.

Why an experiment, not a mandate

This is the core FTTM move applied to the council's own decision: prove the keystone assumption before you scale. The pilot lets an organization find out what AI is worth in its own process at low, contained risk, and it hands the council real evidence instead of a guess. The surveys show both the fearful "no" and the reckless "all-in" backfiring. The pilot is the disciplined middle path.

What the pilot would test

- Planning-cycle time against the current baseline.
- Status-report hours recovered for PMs.
- Rework rate on schedules and decision models.
- PM-reported quality of output and trust in it.
- Any data-handling incident, however minor.

The gate

At week eight the council reviews the measures and makes one of two calls: scale the use and codify the policy from the evidence, or stop and revise. If it stops, it stops on data rather than on fear. Either way the organization ends the window with a defensible policy and no irreversible bets made along the way.

Why now

Every week the work runs without the tools is a week of an aggressive schedule spent slower than it needs to be. The pilot costs the council a measurement window. The full stop costs the program its force multiplier, and on the evidence it does not even buy the safety it promises.

07

Next steps

What we ask of the council

Concretely, lateralworks asks an AI governance council for four decisions.

- **Do not impose a full stop.** Keep the lateralworks tools running on lower-sensitivity data classes while the pilot is scoped.
- **Approve an eight-week pilot** on two named programs, governed by the three-lane framework in Section 04.
- **Agree the success measures and the go/no-go gate up front,** so the week-eight decision is evidence-based rather than contested.
- **Grant a scoped waiver for the most sensitive data** into controlled, non-training (or commercial-terms / isolated) environments, with the council setting the matching control for each class.

lateralworks is not asking any organization to relax its standards. It is asking it to find out what AI is worth before deciding what to allow, and to govern that discovery with a framework that is faster, safer, and more enforceable than a blanket rule. lateralworks will join at whatever depth the council wants for the security discussion, with specialist AI partners available if deeper technical assurance is needed.

The one-line ask. Replace “stop until we have a policy” with “test under controls, measure for eight weeks, then write the policy from what we learn.” It is the safer choice and the faster one.

Sources

References

[1] Stanford Institute for Human-Centered AI. “AI Index Report 2025.” Stanford University, 2025.

[2] Deloitte Global Boardroom Program. “Governance of AI: A critical imperative for today’s boards (2nd edition).” Deloitte, 2025.

[3] S&P Global. “AI adoption is soaring, but few companies are measuring its impact.” S&P Global Sustainable1, October 2025.

[4] Grant Thornton. “2026 AI Impact Survey Report.” Grant Thornton Advisors, 2026.

[5] AvePoint. “The State of AI in 2025.” AvePoint, 2025.

[6] Cato Networks. “Shadow AI Governance Lags as AI Adoption Soars.” Survey of 600+ IT leaders, December 2025.

[7] CIO (Foundry). “Shedding light on shadow AI.” December 2025.

[8] Anthropic. “Updates to our Consumer Terms and Privacy Policy.” Anthropic, 2025–2026. https://www.anthropic.com/news/updates-to-our-consumer-terms

[9] Anthropic Trust Center. Security certifications and reports. https://trust.anthropic.com

[10] Dell’Acqua, F., McFowland III, E., Mollick, E. R., et al. “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.” Harvard Business School Working Paper 24-013, 2023.

[11] Brynjolfsson, E., Li, D., and Raymond, L. R. “Generative AI at Work.” Quarterly Journal of Economics, vol. 140, no. 2, 2025 (NBER Working Paper 31161, 2023).

[12] Noy, S., and Zhang, W. “Experimental evidence on the productivity effects of generative artificial intelligence.” Science, vol. 381, no. 6654, 2023, pp. 187–192.

[13] Juran, J. M. Quality Control Handbook. McGraw-Hill, 1951. Operational source of the 80/20 rule, after Pareto (1896).

[14] Mollick, E. Co-Intelligence: Living and Working with AI. Portfolio, 2024.

[15] lateralworks. “AI-assisted macro planning.” FTTM methodology series, Rev 1.1, 2026. Case-study record of the 80/20 split and the six-station iteration loop.