

Where Should Intelligence Live, and Who Controls It?

Whitepaper

Bart de Graaff · BdG Advisory · October 2026

Every organisation using AI already has an operating model for it, and almost none designed it. Cloud let leaders avoid the question of where intelligence should live and who controls it; AI does not, because every placement of intelligence moves a decision right. This paper shows how ownership, decision rights, risk-tiered governance, incentives, funding and production control fit together as one system.

The operating model you never designed

Every organisation using AI already has an AI operating model. Almost none of them designed it.

It assembled itself. Procurement chose a platform. A business unit funded a pilot. Employees brought their own tools; in Microsoft and LinkedIn's 2024 survey, 78% of AI users at work said they were bringing their own. A risk committee wrote a policy after the first incident. Each of those decisions quietly answered a question nobody had put on the table. Add them up and you have an operating model: a standing answer to where intelligence lives in the organisation and who controls it. It is simply an answer no one chose.

Cloud let organisations avoid that question for a decade. Moving workloads to someone else's data centre changed where computing ran, but not who decided anything. The approval chains, the budget owners, the accountable executives all stayed where they were. AI is different, because it does not only run somewhere. It decides somewhere. Every time a model triages a claim, prices an order, routes a customer or drafts a credit memo, a decision right has moved, and someone either designed that move or inherited it.

That is the structural thesis of this paper. **Operating model design for AI is the deliberate answer to one question, asked decision by decision: where should intelligence live, and who controls it?** The answer has six working parts: ownership, decision rights, governance tiered by risk, incentives, funding and production control. Designed one at a time by different functions, they cancel each other out. Designed as one system, they let an organisation absorb the next capability in weeks instead of years.

The evidence that the operating model, not the model, drives returns is now hard to ignore. McKinsey's March 2025 global survey tested 25 organisational attributes against reported earnings impact from generative AI and found that the redesign of workflows had the biggest

effect. Yet only 21% of respondents using generative AI said their organisation had fundamentally redesigned even some workflows. Most are adding intelligence to a structure built for something else.

This paper sets out how to design that structure: how to place intelligence, who must own it, how to govern it at the right speed, how capital and incentives decide the outcome before anyone notices, and what production control has to look like. Its companion, *Your ERP Is Not Your Company. Your Data Model Is.*, deals with the architecture underneath. This one deals with control.

Placing intelligence is allocating authority

Automation rarely fails because of the technology. It fails because of authority.

The disruption is not the algorithm. It is who no longer makes the decision.

Consider a pattern that repeats across managed services. An operator builds a platform that mines operational events, standardises configurations and deploys agents to correct drift directly on client infrastructure. From an engineering view, it works: security improves, drift disappears, support effort falls. The event mining is welcomed, because it gives senior engineers better information and leaves the decision with them. The automated correction meets a wall, because it executes inside policy boundaries without asking. The outcomes are better and fully compliant. The resistance is about something else. Configuration control was how those engineers held authority, and the system took it.

That is the mechanism. Each placement of intelligence relocates a decision right. If the move is designed, the people who lose authority are told, compensated or redeployed, and the people who gain it are equipped and held to account. If it is not designed, the move happens anyway, by default, and the people who lose authority do the rational thing: they slow it down, route around it or wait for the first error to prove them right.

So the first design act is not choosing a model. It is choosing, for each decision domain, one of four placements:

- **Advise.** The model recommends; a person decides and owns the outcome.
- **Approve.** The model prepares the action; a named person approves each one.
- **Act within bounds.** The model executes inside explicit limits and escalates anything outside them.
- **Act and report.** The model executes; people review patterns, not individual decisions.

The four placements carry very different decision rights. Who approves a new use case? Who approves deployment, a material change to the model, an exception, a shutdown? Most organisations can answer the first question and almost none can answer the last.

The same applies to the decision to automate work at all. In July 2025 Australia's Commonwealth Bank told 45 customer service staff their roles would go, replaced by an AI

voice bot. The Finance Sector Union challenged the decision, saying call volumes were in fact rising and team leaders were being asked to answer calls. In August the bank reversed, conceding that its initial assessment "did not adequately consider all relevant business considerations" and apologising to the staff affected. The voice bot was not the failure. The failure was that the decision about where intelligence would replace people was made on evidence the owners of the work did not accept, without a decision-rights design that would have surfaced the dispute before the announcement rather than after it.

Placement then forces the ownership question, and probabilistic systems make that question harder than it looks. In 2022 a passenger asked Air Canada's website chatbot about bereavement fares and was told he could claim the discount after travel. The airline's written policy said otherwise, and it refused the refund. Before British Columbia's Civil Resolution Tribunal, Air Canada argued, in the tribunal's words, that the chatbot was "a separate legal entity that is responsible for its own actions". The tribunal called this "a remarkable submission", held that the airline is responsible for all the information on its website, and ordered it to pay CA\$812.02.

The sum was trivial. The lesson is not. When something goes wrong, the organisation will be held to own what its AI said and did, whether or not anyone inside it agreed to own it. The only choice is whether ownership is designed in advance or discovered in front of a regulator, a tribunal or a journalist.

Designed ownership has five parts, and they rarely sit with one person by default:

- the **business outcome** the AI is meant to move;
- the **data domain** it draws on;
- the **model's behaviour** in production;
- the **workflow** it changes;
- the **operating risk** it creates.

A regional insurer that automates claims triage typically finds all five owned by different people. Claims owns the outcome. A data team owns the policy records. A platform team owns the model. Operations owns the workflow. Risk owns the risk. When denial rates drift upwards, each owner can show their piece is working. Nobody owns the denial rate.

The design rule is simple to state and hard to enforce: **one named owner of the outcome, with authority over the other four**. The data, model, workflow and risk owners remain accountable for their parts, but they are accountable to the outcome owner for how those parts combine. The NIST AI Risk Management Framework makes the same demand in more neutral language: roles, responsibilities and lines of communication for AI risk should be documented and "clear to individuals and teams throughout the organization". If no one owns the outcome, the exception path and the failure mode, the capability is not ready for production, however well it performed in the pilot.

Govern by risk tier, not by committee

Most organisations govern AI the way they governed software: one approval path, one committee, one set of documents, whatever the use case.

Uniform governance fails in both directions at once. For low-risk uses it is far too slow. A team that wants to summarise internal documents waits in the same queue as a team that wants to automate lending decisions. So people route around it, which is exactly what the bring-your-own-AI numbers describe. For high-risk uses it is too thin, because a committee that reviews everything can examine nothing in depth. The organisation ends up with heavy process and weak control: the worst of both.

The mechanism behind this is rational on every side. The committee is accountable for incidents, not for value forgone, so it optimises for refusal and delay. Business teams are accountable for results, not for risk outside their function, so they optimise for speed and for staying off the committee's agenda. Both behave sensibly. The system they create is unsafe and slow.

The way out is not to be found in AI ethics literature. It is in regulators who have done this before. Banking supervisors have governed models for fifteen years. The US Federal Reserve's model risk guidance, SR 11-7, was replaced in April 2026 by SR 26-2, which asks for a risk-based approach "tailored to a banking organization's model risk profile": the depth of control follows the risk of the model, not a uniform checklist. The EU AI Act takes the same shape at legislative scale. It sorts systems into four levels: unacceptable, high, transparency and minimal risk, and reserves its heavy obligations for the high-risk tier, most of which now apply from December 2027.

Tiering is therefore not a concession to speed. It is how mature control systems already work. An enterprise version needs three or four tiers, set by two questions: how consequential is a wrong decision, and how reversible is it?

- **Tier 1: low consequence, easily reversed.** Internal drafting, search, summarisation. Pre-approved platforms, standard data rules, light logging. Teams self-certify and move.
- **Tier 2: material but reversible.** Customer-facing content, operational recommendations, pricing within bands. A named owner, a documented design review, defined monitoring, a clear escalation path.
- **Tier 3: high consequence or hard to reverse.** Credit, claims denial, safety, employment, anything regulated. Independent validation, human authority over individual decisions or tight bounds, continuous monitoring and a rehearsed shutdown.

Each tier gets its own approver, its own evidence and its own speed. The committee stops reviewing everything and starts setting the tier boundaries and auditing the tiers. That is the shift from governance as approval to governance as design: governance shapes how AI operates from the start rather than appearing at the end as the last gate.

The most useful governance design for Tier 2 replaces decision approval with outcome boundaries. A retailer that lets store-level systems set prices does not need a regional

manager to approve each price. It needs explicit bounds on margin and on price relative to competitors, automatic escalation outside them, and someone who reviews the pattern of decisions each week. Oversight moves from the individual decision, where people are slow, to the boundary and the pattern, where people add judgement. Governance runs at the speed of the system without abandoning control.

Capital decides where intelligence lives

Executives often describe AI placement as a technical decision. In practice it is made by the budget and the scorecard, long before an architect is consulted.

The structural loop that drives execution runs through capital, not strategy. Capital concentrates control, because whoever funds the work decides what it is for. Control shapes what data is collected and which metrics are reported. Data drives execution, because teams optimise for the numbers they see rather than the strategy they are told. Execution produces outcomes, and the outcomes decide where capital flows next. AI does not change the loop. It accelerates every turn of it, which means a misaligned loop now produces misaligned results in weeks rather than years.

Klarna is the clearest public example of the loop at work. In February 2024 the company announced that its AI assistant had handled 2.3 million conversations in its first month, two-thirds of its customer service chats, equivalent to the work of 700 full-time agents, and estimated it would drive a US\$40 million profit improvement that year. Fifteen months later its chief executive told Bloomberg that the company would invest in human support again, because cost had been "a too predominant evaluation factor" and the result was lower quality.

Read structurally, the lesson is not that AI cannot serve customers. It is that the evaluation metric decided where intelligence lived. When cost is the dominant measure, the operating model places intelligence wherever it removes cost, and quality is discovered as a casualty later. The placement was an incentive decision presented as a technology decision.

Incentives work the same way inside the organisation. A sales team paid on this quarter's bookings will ignore a model that recommends lifetime value. A plant manager measured on maintenance cost will defer the model's early intervention. A partner whose standing rests on staffing authority will resist an engine that allocates people firm-wide. None of these people are wrong by the logic of their scorecard. The operating model is wrong for leaving the scorecard untouched while moving the decision.

Funding has its own version of the loop. Most AI budgets are pilot budgets: short, project-shaped and owned by whichever function sponsored the experiment. They pay for proving that the model works. They rarely pay for what production needs: the data work, the monitoring, the controls, the support rota, the retraining of the people whose work changes. Gartner predicts that over 40% of agentic AI projects will be cancelled by the end of 2027 because of escalating costs, unclear business value or inadequate risk controls. Read those three causes again. Not one of them is a model failure. Each is an operating model failure: a

funding model that did not price production, an incentive model that did not define value, and a governance model that did not build controls.

The design response is to fund the operating system rather than the use case. That means standing budgets for the shared platform, data readiness, governance and production support, with use cases drawing on them; a funding gate between pilot and production that requires a named outcome owner and a production cost estimate; and metrics that track value captured in the business, not models deployed or users onboarded. It also means changing the scorecards of the people whose authority moves, in the same budget cycle as the deployment, not two years later. McKinsey's survey points the same way: CEO oversight of AI governance was among the attributes most correlated with reported bottom-line impact. The person who owns the capital loop has to own the design.

Production control is the operating model at runtime

Pilots need speed. Production needs control. Most organisations use the same team, the same budget and the same governance for both, then wonder why pilots do not survive contact with the business.

Production control is where every other design decision is either enforced or exposed. Ownership is meaningless if the owner never sees the metrics. Tiering is meaningless if a Tier 3 system has no tested way to be stopped. Decision rights are meaningless if nobody has pre-agreed authority to pull the system back when it drifts.

Zillow Offers shows what that costs at scale. Zillow bought homes using its own price forecasts and planned to resell them. In November 2021 it announced the wind-down of the business, a write-down of inventory of approximately US\$304 million for homes bought above what it now expected to sell them for, a further expected loss of US\$240–265 million in the following quarter, and a reduction of about a quarter of its workforce. Its chief executive said that "the unpredictability in forecasting home prices far exceeds what we anticipated."

Forecast error in a volatile market is not a surprise; it is the known risk of the business. The structural question is what the operating model did with it. Whatever the internal details, the public numbers show the decision to keep buying at scale outran the evidence that the forecasts were holding. That is a production control problem: the link between the model's live performance and the authority to throttle the business it drives.

Good production control is specified before go-live, in writing, for every Tier 2 and Tier 3 system:

- **Monitored measures** tied to the business outcome, not only to model accuracy, reported to the outcome owner.
- **Thresholds** that trigger action automatically: slow down, narrow the bounds, revert to human decision.

- **Escalation paths** with named people and response times, tested before they are needed.
- **Intervention and shutdown authority** held by a named person who does not need a committee to act.
- **Review cadence** for model changes, data changes and drift, scaled to the tier.
- **Decommissioning**, because systems that cannot be retired safely become permanent by accident.

None of this is novel. The NIST framework asks for mechanisms, with responsibilities assigned and understood, to "supersede, disengage, or deactivate" AI systems that perform outside their intended use, and for processes to decommission them safely. What is unusual is an organisation that has done it for every consequential system before the first incident rather than after.

Production control is also where the operating model learns. The monitoring that tells you when to stop a system also tells you where the boundaries could safely widen, which tier a use case really belongs in, and which owners are carrying more than they can see. An organisation that treats production control as compliance overhead loses that feedback. One that treats it as the runtime of its operating model gets better at placing intelligence every quarter.

The strongest objection: design will kill the experiment

The best argument against all of this deserves a fair hearing.

It runs as follows. AI capability changes every few months. Any operating model designed today will describe a world that no longer exists by the time it is approved. The organisations that have learned fastest did so by letting people use the tools, watching what worked and codifying it afterwards. Heavy design up front is how large companies turn a fast-moving technology into a slow-moving programme. The enterprise that spends a year designing decision rights will find its competitors already in production, and its own staff using consumer tools anyway. Better to let the operating model emerge from use.

Much of this is right. Experimentation should be fast, cheap and lightly governed. Designing the AI equivalent of a five-year ERP blueprint would be a mistake. And bottom-up use does reveal where value really sits, often in places no central team would have predicted.

But the objection confuses two different activities. Experimentation and production are not the same thing, and they should not share an operating model. The design work this paper describes is aimed at the boundary between them and at what lies on the production side of it: who owns an outcome once it touches customers or money, which tier it sits in, what it is funded from and how it is stopped. That boundary is exactly what lets experimentation stay fast. Teams can try anything in Tier 1 because everyone knows what it takes to cross into Tier 2.

The objection also assumes that not designing is neutral. It is not. An undesigned operating model is still an operating model, with its parameters set by whoever acts first: the procurement team's platform choice, the finance team's cost metric, the first incident's policy response. Klarna did not lack experimentation. It lacked a design that put quality, not only cost, into the decision about where intelligence would live.

Finally, the design does not need to chase capability, because it is built around decisions, not tools. A claims decision, a pricing decision or a credit decision keeps the same consequence and the same reversibility whichever model is making it. When capability improves, the right response is to move a decision domain from one placement to another, from approve to act within bounds, with its owner, tier and controls already in place. That is a quarterly adjustment, not a redesign. The organisations that move fastest are the ones that have already answered the authority question, which is the argument of the sibling paper on governance-first deployment.

Structural moves, in order

The work is not to add AI to the current system. It is to redesign the system so that AI can operate safely, repeatedly and economically. The order matters, because each step depends on the one before.

First, map decisions, not use cases. List the twenty or thirty decision domains where AI will matter most: credit limits, claims, pricing, scheduling, routing, hiring screens. For each, answer the paper's question explicitly. Where should intelligence live: advise, approve, act within bounds, or act and report? Who controls it? Write the answer down. Most organisations discover that nobody has ever stated it.

Second, name the outcome owner for each domain, with authority over the data, model, workflow and risk owners who serve it. If no executive will accept the outcome, the domain is not ready for anything beyond Tier 1.

Third, set the risk tiers and give each its own path. Define the tiers by consequence and reversibility, assign every domain to one, and give each tier its own approver, evidence, monitoring and speed. Move the central committee from reviewing cases to setting boundaries and auditing them.

Fourth, rewire capital and incentives in the same cycle. Fund the shared platform, data readiness, governance and production support as standing capabilities. Put a gate between pilot and production that requires an owner and a production cost. Change the scorecards of the people whose authority moves, at the same time as the deployment.

Fifth, specify production control before go-live. Measures, thresholds, escalation, shutdown authority, review cadence and decommissioning, written down and tested for every Tier 2 and Tier 3 system.

Sixth, review the operating model as one system. Once a quarter, the outcome owners, the risk function and the capital allocator review placements, tiers, funding and incidents together, and move decision domains between placements as evidence accumulates. This is where the operating model compounds.

Leaders should also stop doing four things:

- **Stop funding pilots that have no path to production,** no named owner and no production cost estimate.
- **Stop routing every AI use through one committee.** It makes low-risk work slow and high-risk work thinly governed.
- **Stop measuring adoption.** Users onboarded and models deployed are activity. Value captured is the measure.
- **Stop letting procurement and cost metrics decide placement.** Where intelligence lives is an authority decision and belongs to the people accountable for the outcome.

Two organisations can license the same model on the same day. One deploys it into a structure that routes every consequential decision through a committee that meets fortnightly. The other has already decided, domain by domain, where intelligence lives, who owns it, how fast it may act and who can stop it. Same technology, different outcome, because the constraint was never the model.

Intelligence will end up living somewhere in your organisation; the only question is whether you decided where.

What to do about it

- **Map decisions, not use cases.** For the twenty or thirty decision domains that matter most, write down where intelligence should live (advise, approve, act within bounds, act and report) and who controls it.
- **Name the outcome owner.** Give one executive each outcome, with authority over the data, model, workflow and risk owners who serve it. No owner, no production.
- **Tier governance by consequence and reversibility.** Give each tier its own approver, evidence, monitoring and speed, and move the central committee from reviewing cases to setting and auditing the tiers.
- **Rewire capital and incentives in the same cycle.** Fund platform, data readiness, governance and production support as standing capabilities, gate pilots on a named owner and a production cost, and change scorecards when the decision moves.
- **Specify production control before go-live.** Write down measures, thresholds, escalation, shutdown authority, review cadence and decommissioning for every material system, and test them.
- **Review the operating model as one system.** Each quarter, review placements, tiers, funding and incidents together, and move decision domains between placements as the

evidence accumulates.

Sources

- Microsoft and LinkedIn, 2024 Work Trend Index: AI at Work Is Here, 2024: <https://news.microsoft.com/source/2024/05/08/microsoft-and-linkedin-release-the-2024-work-trend-index-on-the-state-of-ai-at-work/>
- McKinsey, The state of AI: How organizations are rewiring to capture value, March 2025: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-how-organizations-are-rewiring-to-capture-value>
- Information Age (ACS), CBA reverses AI-driven job cuts, admits 'error', 2025: <https://ia.acs.org.au/article/2025/cba-reverses-ai-driven-job-cuts--admits--error-.html>
- McCarthy Tétrault, Moffatt v. Air Canada: a misrepresentation by an AI chatbot, 2024: <https://www.mccarthy.ca/en/insights/blogs/techlex/moffatt-v-air-canada-misrepresentation-ai-chatbot>
- The Register, Air Canada must honour refund policy invented by airline's chatbot, 2024: https://www.theregister.com/2024/02/15/air_canada_chatbot_fine/
- NIST, AI Risk Management Framework core (GOVERN 2.1, GOVERN 1.7, MANAGE 2.4), 2023: <https://airc.nist.gov/airmf-resources/airmf/5-sec-core/>
- Board of Governors of the Federal Reserve System, SR 26-2: Revised Guidance on Model Risk Management, 2026: <https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm>
- European Commission, AI Act: regulatory framework for AI: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
- Klarna, Klarna AI assistant handles two-thirds of customer service chats in its first month, 2024: <https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/>
- Entrepreneur, Klarna is hiring customer service agents after AI couldn't cut it (CEO interview with Bloomberg), 2025: <https://www.entrepreneur.com/business-news/klarna-ceo-reverses-course-by-hiring-more-humans-not-ai/491396>
- Gartner, Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027, 2025: <https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027>
- Zillow Group, Third quarter 2021 results; plan to wind down Zillow Offers, 2021: <https://investors.zillowgroup.com/news-and-events/news/news-details/2021/Zillow-Group-Reports-Third-Quarter-2021-Financial-Results--Shares-Plan-to-Wind-Down-Zillow-Offers-Operations/default.aspx>

Read online: bdgadvisory.com/perspectives/operating-model-design