

The Last 20%: Margin Is Not Reinvention

Whitepaper

Bart de Graaff · BdG Advisory · October 2026

AI business cases promise a margin and are presented as transformation. But when systems take the routine volume, what remains is harder, riskier and organised for a mix that no longer exists. This paper explains why capability outruns absorption, why the last 20 percent is a different job, and what to build before automating the first 80.

Margin is not reinvention

Every AI business case promises a margin. Almost none promise a different business.

That is the innovation gap.

When AI takes over most of the volume in a function, what remains is not a smaller version of the old function. It is a different job. The routine work leaves first. The exceptions, the judgement calls, the angry customers, the disputed claims and the decisions somebody must answer for stay behind. That residual work needs a different structure: different roles, different authority, different metrics, a different cost model and a different way of growing the people who do it. Most organisations bank the margin and keep the structure. They optimise the existing model while the architecture underneath it shifts, and each round of optimisation compounds the wrong thing.

Call it the last 20 percent. The exact split varies by function and by year. The shape does not: a large share of volume handled by systems, and a smaller, harder share left to people who were organised for the old mix.

The distinction between margin and reinvention is worth making precise. Margin improvement does the same work, for the same customers, through the same structure, at lower cost. Reinvention changes at least one of those three: the work, the customer or the structure. Both are legitimate. The problem is that AI business cases are almost always written as the first and then presented as the second. The board hears "transformation". The plan delivers a cheaper version of today.

The operating model gap describes why AI so often fails to get into operation at all. This paper starts where that one ends, with AI that works. It explains why organisations cannot absorb capability as fast as it arrives, why the residual work needs its own design, why the margin case wins the budget anyway, and what leaders should build for the last 20 percent before they automate the first 80.

Capability arrives faster than organisations can absorb it

The gap between invention and impact is old, and it is structural.

Electricity is the clearest case. As Tim Harford summarised the economic historian Paul David's work, electric light and generating stations existed by the early 1880s, yet factories that replaced their steam engine with an electric dynamo saw productivity barely improve. Steam-era factories were built around a single engine driving every machine through shafts and belts. Swapping the engine kept the layout, the workflow and the logic of the building intact. The gains came decades later, in the 1920s, when factories were redesigned on a single floor around the flow of work, with each worker running their own electric motor. The technology had been available for forty years. The structure took that long to change.

Modern economics has a name for the pattern. Erik Brynjolfsson, Daniel Rock and Chad Syverson describe a productivity J-curve: general-purpose technologies such as AI require large complementary investments in new processes, products, business models and human capital. Those investments are mostly intangible and poorly measured, so early on the technology looks less productive than it is, and the payoff arrives only once the complementary work is done.

That complementary work is what this paper calls absorption. Absorption capacity is the rate at which an organisation can redesign itself around a new capability. It has several components, and each moves at its own speed:

- **Roles:** what people are hired, trained and promoted to do.
- **Decision rights:** who may decide what, and at which limit.
- **Data:** what is captured, who owns it and who may use it.
- **Metrics:** what counts as good performance, and for whom.
- **Economics:** how the function is costed, priced and funded.
- **Talent pipeline:** how today's juniors become tomorrow's experts.

Absorption is set by the slowest of these, not the fastest. A model can be deployed in weeks. A role redesign takes a year of consultation. A change to the cost model waits for the budget cycle. A talent pipeline takes five years to show whether it worked. Capability arrives at software speed and is absorbed at organisational speed. The difference between the two is the innovation gap, and it widens every time capability improves.

The current wave shows the strain. IBM's 2025 CEO study, a survey of 2,000 chief executives, found that only 25 percent of AI initiatives had delivered expected ROI over the previous few years and only 16 percent had scaled enterprise-wide. Half of the executives said rapid investment had left them with disconnected technology. That is what capability without absorption looks like from the top: a growing estate of tools and a structure that has not caught up.

Consider a mid-sized accounting firm. In eighteen months it deploys AI for document extraction, reconciliation, first-draft tax workpapers and client correspondence. Each tool

works. The firm's structure is still the pyramid it was built on: graduates doing preparation, seniors reviewing, managers managing clients, partners signing. Staffing ratios, billing rates and promotion timelines all assume that pyramid. The tools have removed much of the work at the bottom, and the firm has not decided what the bottom is now for. It has the dynamo. It still has the steam-era building.

The last 20 percent is a different job

AI does not take a random slice of the work. It takes the easiest, most repetitive, highest-volume slice first. What remains is concentrated difficulty.

Klarna is the public test case. 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, doing the equivalent work of 700 full-time agents and cutting resolution time from 11 minutes to under two. It estimated a \$40 million profit improvement for 2024. By May 2025 the chief executive was describing a correction. Cost, he told Bloomberg, had been "a too predominant evaluation factor", and the result was lower quality, as reported by CX Today. Klarna began recruiting human agents again under a flexible, Uber-style model.

Read the sequence structurally. The AI did what it was designed to do with the volume. The company then had to design, under pressure, a new model for the remainder: who handles it, how they are engaged, what quality means for the cases that remain. That design work was not in the original business case. It became necessary the moment the business case succeeded.

Australia produced a sharper version. In July 2025 the Commonwealth Bank announced that 45 customer service roles would be replaced by an AI voice bot. The Finance Sector Union said calls were in fact increasing and that management was offering overtime and asking team leaders to answer calls. In August the bank reversed the decision, saying its initial assessment "did not adequately consider all relevant business considerations" and that the roles were not redundant. By its own account, the bank's assessment had missed relevant considerations. The union's account suggests which: the work that remained after the bot took its share.

The mechanism is the same in every function. Operations are built for a blended mix of easy and hard work. Staffing models, handling-time targets, training, career paths, spans of control and quality measures all assume that blend. Remove the easy volume and every one of those averages becomes misleading. The remaining cases take longer, need more authority to resolve and carry more risk. A team measured on the old targets looks worse at exactly the moment it is doing harder and more valuable work.

The customer side changes too. In a blended operation, a customer reaching a person was usually at the start of their problem. In an automated one, the customer reaching a person has typically already failed with the system. They arrive later, more frustrated, often having repeated themselves, and with a problem the system could not classify. The handoff between system and person becomes one of the most important moments in the customer

relationship, and in most organisations nobody designed it. It was left as an escalation path, built for the exception rather than for what is now the main event.

The economics flip as well. Total cost falls, but the cost per remaining case rises, because every remaining case is a hard one. A finance team watching cost per contact or cost per claim sees the number go up after automation and concludes the people are underperforming. The pressure that follows is to squeeze the residual team on the old unit-cost measures, which is how quality collapses in the cases that matter most to customers and regulators. The unit of account has to change with the work. If it does not, the management system will punish the residual team for the success of the automation.

There is a second change, less visible and more important. The people in the residual role must now judge the system as well as the case. A field experiment with 758 Boston Consulting Group consultants, led by Harvard Business School researchers, makes the point. On tasks within the model's capability, consultants using AI completed more tasks, faster and to higher quality. On a task deliberately chosen to sit outside it, the control group answered correctly 84.5 percent of the time; consultants with AI did markedly worse, between 60 and about 71 percent, because they trusted the analysis and followed it to the wrong conclusion. The researchers called this the jagged frontier.

The last 20 percent sits on that frontier. Knowing when the system is wrong, overriding it and owning the consequence is a new skill and a new accountability. It does not exist in the old job description, and it cannot be bolted onto a role designed for throughput.

Picture a general insurer's claims function. AI now settles straightforward motor claims within hours. The adjusters who remain handle disputed liability, suspected fraud, injury and vulnerable customers. They are still measured on claims closed per day, still supervised in teams of fifteen, still escalating anything above a modest limit to a manager. Every part of that structure was right for the old mix. Every part is wrong for the new one.

Optimisation compounds the wrong thing

Optimisation feels safe because each step pays for itself. That is exactly why it is dangerous when the architecture is shifting.

Every round of optimisation makes the existing model cheaper and more entrenched. Processes are tuned, contracts renewed, metrics refined, systems integrated more tightly. The organisation becomes very good at a model whose foundations are moving. The investment is real, and so is the lock-in.

Containerisation shows how this ends. Containers needed deep-water ports and integrated road and rail networks. As LSE research on British shipping notes, many older ports, such as the Port of London, were unsuited to the new technology and declined, while new ports expanded in more suitable locations such as Felixstowe, and port activity concentrated in a few places. Getting better at handling cargo the old way did not help. The unit of competition had changed from the efficiency of the dock to the design of the network.

Outsourcing shows the same trap at enterprise scale. Many back-office functions are already run through business process outsourcing contracts priced per transaction or per seat. The MIT NANDA study of enterprise AI pilots, as reported by Fortune, found the biggest returns in back-office automation, including eliminating business process outsourcing and cutting external agency costs. The optimising response is to use AI to renegotiate the outsourcing rate. The reinventing response is to ask whether the function, designed around the transaction volumes that justified the contract, should exist in its current form at all. The first produces a better contract for a model in decline. The second produces a decision.

AI is doing the same to many functions. The data-driven architecture perspective argues that processes are becoming temporary vehicles for decisions rather than core infrastructure. An organisation that spends three years optimising a process with AI may be perfecting something the architecture is about to make disposable.

The most serious compounding happens in people. AI takes the junior work first. Stanford's Digital Economy Lab, using payroll data covering millions of US workers, reports in its August 2026 revision that employment of workers aged 22 to 25 in AI-exposed occupations now stands 19 percent below where it would be had it kept pace with less-exposed peers, with no comparable gap for experienced workers. The adjustment is happening mainly through reduced hiring, and it is concentrated in occupations where AI substitutes for tasks rather than complementing them.

Now connect that to where AI's gains come from. In a large field study of generative AI at work, covering 5,179 customer support agents, access to an AI assistant raised productivity by 14 percent on average and by 34 percent for novice and low-skilled workers, with minimal impact on the most experienced. The authors found suggestive evidence that the system spreads the know-how of the most able workers to newer ones.

Put the two findings side by side. The system is valuable because it carries the judgement of experienced people. The work that used to turn juniors into experienced people is the work the system now does. An organisation that optimises the 80 percent by cutting graduate intake is consuming the pipeline that produces the people the 20 percent depends on. It will not notice for five years. Then it will have a capable system and nobody senior enough to know when it is wrong.

That is what compounding the wrong thing means. Each year's saving is real. The structural debt grows underneath it.

Why the margin case always wins the budget

If the residual structure matters so much, why do organisations keep funding the margin and skipping the redesign? Because every part of the budgeting system favours it.

A margin case is measurable. It fits inside one function, one cost centre and one budget year. Finance can validate it, the function head can deliver it, and investors understand it. Klarna's announcement led with an estimated profit improvement for a single year. That is

how AI cases are written, because that is how they get approved.

A reinvention case is the opposite. It crosses functions, so it has no natural owner. Its benefits arrive later than its costs. It asks the people who run the current model to design their own replacement, which is the problem the structural transformation perspective addresses directly: legacy governance cannot reform itself from inside. And it cannot be validated against last year's numbers, because it changes what the numbers measure.

Careers reinforce the bias. The executive who delivers a 20 percent cost reduction in her function has a result to show at the next review. The executive who spends the same year redesigning what the function is for has a plan, a set of difficult conversations and a cost line that went up. Promotion systems reward the first. Over a few cycles, the leadership bench fills with people who are excellent at optimising and have never been asked to redesign anything.

So the system selects for optimisation. Function heads propose the savings they can control. Finance banks them. The saving shows up as a smaller cost base, which is then defended as the new baseline. Nobody is accountable for what the function should become, because nobody was asked.

Organisations that focus primarily on cost reduction misread the shift. They optimise within an existing model while the structure changes around them. The misalignment surfaces later as execution failure, and by then it is usually blamed on the technology.

Consider a national telco's enterprise division. AI handles quote generation, contract drafting and first-line technical support. The division books the saving and reduces its sales support headcount. Meanwhile customers are buying connectivity bundled with security and cloud services, a sale that needs fewer people but very different ones: solution architects, partner managers, people able to price risk. The savings funded a leaner version of the division that was selling the old way. Nobody had been asked to design the division that sells the new way.

The strongest objection: optimise first, reinvent later

The best counter-argument is serious and should be stated fairly.

Optimisation funds reinvention. Margin today pays for experiments tomorrow. Grand reinvention programmes have a poor record: they consume years and capital, and the design is often wrong because the technology moves faster than the programme. AI capability is still unstable, and reorganising around today's models risks building for a frontier that shifts next year. Incremental improvement lets an organisation learn what the technology can really do before betting the structure on it. Sequence matters: optimise first, reinvent once the picture is clear.

Much of this is right. Big-bang redesigns usually fail. Learning by deploying is better than designing in the abstract. Margin is a legitimate source of funding.

The flaw is in the word "later", for three reasons.

First, optimisation without residual design does not even deliver the margin. Klarna and the Commonwealth Bank were not reinvention programmes. They were optimisation programmes, and both had to be partly reversed. In both, the public explanation pointed at the work that remained: quality in one case, call demand in the other. The residual structure is not a second phase. It is a condition for the first phase to hold.

Second, "later" has deadlines that do not wait for clarity. The talent pipeline is one: graduates not hired this year are the experienced staff missing in 2031. The architecture is another: each year spent optimising a process deepens the investment in something that may become disposable.

Third, the funding logic only works if it is designed. Savings that are banked as a smaller baseline do not fund anything. If optimisation is meant to pay for reinvention, a defined share of the saving has to be ring-fenced for it, with an owner, from the start.

So the answer is not reinvention instead of optimisation. It is optimisation designed as the first move of a reinvention: automate the volume, and in the same decision design the residual structure and commit part of the margin to building it.

Designing for the last 20 percent

The structural moves follow from the argument, and their order matters.

First, map the residual before you automate the volume. For every function where AI will take significant volume, describe what will remain: the case mix, the complexity, the risk, the authority it needs. If the business case cannot describe the residual work, it is not finished.

Second, redesign the residual role around judgement and authority. Give the people who handle the remaining cases the authority to resolve them, not just route them. Make frontier judgement explicit: knowing when to override the system is part of the job, and so is owning the outcome. Write it into role design, training and accountability.

Third, re-base the metrics and staffing for the new mix. Retire averages built for the blended workload. Measure resolution quality, risk outcomes and customer value for the residual cases. Re-plan spans of control, escalation limits and staffing levels for harder work.

Fourth, ring-fence part of the margin. Before any saving is banked, commit a defined share to building the residual structure and the next model. Give that money an owner who is not the head of the function being optimised.

Fifth, protect the pipeline that makes experts. Decide deliberately how people will build judgement when the system does the work that used to teach it. Apprenticeship, rotation and supervised exposure to hard cases have to be designed rather than left to happen, because the old path to expertise is the one AI removed.

Sixth, ask whether the function is still the right unit. When the architecture shifts, the boundaries of functions often stop making sense. Review whether the residual work belongs together, or whether it now sits more naturally with the decision it serves.

The order is deliberate. The residual map tells you what the new role must do. The role tells you what to measure. The measures tell you what the new structure costs, which tells you how much of the margin to ring-fence. The pipeline question only becomes answerable once you know what expertise the residual role needs. And the question of whether the function survives is last because it is the hardest, and because it should be answered with evidence from the first five moves rather than from a strategy offsite.

Return to the insurer's claims function and run it the other way. Before the motor-claims automation goes live, the claims director maps what will remain: disputed liability, fraud, injury, vulnerable customers. She redesigns the adjuster role around those cases, with settlement authority raised so adjusters resolve rather than escalate, and with explicit responsibility for overriding the system when its assessment looks wrong. Claims closed per day is replaced by resolution quality, leakage and complaint outcomes. Teams shrink from fifteen to eight per supervisor because the work is harder. A fixed share of the automation saving funds a claims academy where new adjusters work supervised cases the system has flagged as difficult, so the next generation of senior adjusters is built on purpose rather than by accident. The automation and the first-year saving are identical. Only the second insurer ends up with a claims function that fits what claims have become.

Leaders should also stop four habits. Stop counting headcount reduction as transformation. Stop banking the full saving as next year's baseline. Stop measuring residual teams on targets built for the old mix. And stop cutting graduate intake before deciding how the organisation will make its future experts.

None of this slows AI adoption. It makes the adoption hold. The operating model gap explains why AI fails to get into operation. The innovation gap explains why, once it does, the organisation can still end up as a cheaper version of a business that no longer fits the architecture.

AI takes the volume. What remains is the business.

What to do about it

- **Map the residual first.** Before automating volume, describe the case mix, complexity, risk and authority of the work that will remain. A business case without that description is unfinished.
- **Redesign the role around judgement.** Give residual teams authority to resolve rather than route, and make overriding the system and owning the outcome an explicit part of the job.
- **Re-base metrics and staffing.** Replace averages built for the blended workload with measures of resolution quality, risk and customer value, and re-plan spans and escalation limits for harder work.

- **Ring-fence part of the margin.** Commit a defined share of every AI saving to building the residual structure and the next model, owned by someone other than the head of the function being optimised.
- **Protect the pipeline that makes experts.** Design apprenticeship, rotation and supervised exposure to hard cases deliberately, because AI removes the work that used to build judgement.
- **Ask whether the function is still the right unit.** Once the residual is clear, test whether the remaining work belongs together or now belongs with the decision it serves.

Sources

- Tim Harford, Paul David: It took decades for the economic impact of electricity to become clear (Financial Times, 2007, via History News Network): <https://www.historynewsnetwork.org/article/paul-david-it-took-decades-for-the-economic-impact>
- Brynjolfsson, Rock and Syverson, The Productivity J-Curve, NBER Working Paper 25148: <https://www.nber.org/papers/w25148>
- IBM Institute for Business Value, 2025 CEO Study press release, May 2025: <https://newsroom.ibm.com/2025-05-06-ibm-study-ceos-double-down-on-ai-while-navigating-enterprise-hurdles>
- Klarna, AI assistant handles two-thirds of customer service chats in its first month, February 2024: <https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/>
- CX Today, Klarna grapples with AI-led customer service, pivots to an Uber-type of setup, May 2025: <https://www.cxtoday.com/contact-center/klarna-grapples-with-ai-led-customer-service-pivots-to-an-uber-type-of-setup/>
- ACS Information Age, CBA reverses AI-driven job cuts, admits error, August 2025: <https://ia.acs.org.au/article/2025/cba-reverses-ai-driven-job-cuts--admits--error-.html>
- Harvard Business School AI Institute, summary of the jagged technological frontier field experiment with BCG consultants: <https://aiinstitute.hbs.edu/back-to-the-beginnings-of-ai-at-work/>
- LSE Business Review, Dockworkers and the introduction of containers in UK shipping in the late 1960s, 2017: <https://blogs.lse.ac.uk/businessreview/2017/11/10/dockworkers-and-the-introduction-of-containers-in-uk-shipping-in-the-late-1960s/>
- Fortune, MIT report: 95% of generative AI pilots at companies are failing, August 2025: <https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/>
- Stanford Digital Economy Lab, Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence, revised August 2026: <https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/>

-
- Brynjolfsson, Li and Raymond, Generative AI at Work, NBER Working Paper 31161:
<https://www.nber.org/papers/w31161>

Read online: bdgadvisory.com/perspectives/innovation-gap