

A UST EVOLVE POINT OF VIEW

The retail cost-to-serve reset

Where the next productivity advantage will be won, and why it is bigger than automation.



WHITEPAPER

UST Evolve
Retail Transformation

September 2026

ust.com

Decision latency is the diagnosis.
Decision compression is the response.
A structurally cheaper business is the outcome.

TABLE OF CONTENTS

Inside this point of view

Decision latency → decision compression → structural productivity

Where this started	1
The diagnosis	2
The market in numbers	3
The productivity paradox	4
Why now	5
Where the cost actually comes from	6
How latency compounds	7
Patterns	10
The conversion problem	12
Evidence	13
What we are doing about it	14
Funding	15
The other half of the argument	16
Getting started	17
A missing discipline	20
Ownership and governance	21
Foundations	23
The last mile	24
A practical path to scale	25
The scorecard	26
The end state	27
Final thoughts	28



WHERE THIS STARTED

Not long ago, I sat down with an executive team that knew exactly what it wanted, which, in this industry, is rarer than it ought to be. It was an RFP; we were one of several firms invited to pitch, and the brief's objective was headcount. Reduce it.

I wasn't first through the door. The usual consulting suspects had already sat in the very chair I was now in, and each had answered the mandate as written: cut labor head by head, redraw the org chart, flatten a reporting line, hunt down duplicate roles, benchmark spans against peers. Reasonable work, competently done, I'm sure. It is also what you get when several firms are handed the same question and every one of them decides the safest way to win is to agree with it.

Then it was my turn, and we did the ten minutes everybody does. How were the weekends? How about this heatwave? Good to see you again, Nancy. Then I took the spotlight.

"You do not have a headcount problem. You have a decisioning problem."

I'll be honest with you. Judging by the faces in my audience, it didn't land well. You could feel the room decide I'd misread the brief, which is hard to watch when you are the last item on the agenda and the brief is why you were invited. But it bought me the first of the two things I wanted - somebody at the far end of the table asking, "What does that mean, Greg?"

This paper is the long answer.

THE DIAGNOSIS

Retail thinks it has a headcount problem. What it actually has is a decision latency problem.

Retail does have a cost problem. That part isn't in dispute. What's in dispute is where it comes from, and that distinction matters more than it sounds. For thirty years the industry has gone after operating costs the same way: restructure, outsource, build a shared service center, modernize the platform, automate what can be automated, and when times get tough, cull the herd; plenty of that created real value. However, all of that shares one blind spot: it starts with the cost you see on a payroll report rather than the work that creates the cost in the first place. The result is a smaller payroll stapled to an unchanged process.

Headcount in most retailers is just a symptom, not the disease. Look at where the hours go, and you find something more specific. Decisions that take three weeks when the market moved in three days. The same forecast is rebuilt by merchandising, supply chain, procurement, and finance while they all play tug-of-war with each other. Analysts spent Monday and Tuesday assembling the data for Wednesday's meeting, which needs to reach a conclusion most people in the room already suspected.

Here's a simple way to think about this: when heads get cut, and work gets redistributed, that work doesn't go away. It's the same amount of work just redistributed to fewer people. Does this reduce actual work that needs to be done? Does this accelerate any timelines? Does this cause higher employee turnover? I know the answers. You know the answers.

That is why the real prize here is bigger than automation and different from it. Call it 'decision compression': cutting the elapsed time, handoffs, analysis, approvals, reconciliations, and the sheer human horsepower required to get from a signal to an action. Automation asks how to do a task faster. Decision compression asks whether the task needs to exist at all.

If we're going to keep leaning on the term, it needs a number attached. Decision latency is signal-to-action time: the elapsed days between when the data says something has changed and when a customer can see us respond. Measure it per decision type, not as a company average, because a markdown and a supplier commitment are not the same animal. Most retailers have never measured it at all. The ones who have are usually unhappy with the answer.

For any retailer writing large AI checks, that difference is the whole game. The money isn't in layering some form of intelligence over everything, but in using that intelligence to change how the business runs. When latency drops, a surprising amount of secondary work evaporates, because most of it existed to cope with the delay. When functions consider a decision together instead of passing it down a line, duplicate versions stop being produced. When routine execution moves into a governed workflow, capacity comes back, and only then does the cost base have anywhere to go.

Automation asks how to do a task faster.
Decision compression asks whether the
task needs to exist at all.

THE MARKET IN NUMBERS

The industry has bought the intelligence. It has not bought the execution.

Start with the size of what is still sitting on the table.

IHL Group estimates inventory distortion, the combined cost of empty shelves and unsold stock, at roughly \$415 billion a year across North America. Globally, the number is about \$1.7 trillion, or 6.2% of retail sales, down from 10.4% in 2021. So, five years of forecasting, inventory systems, data platforms, and automation clearly moved it. What has not moved much is the residue: six bucks in every hundred dollars of sales, still leaking out.

The forward-looking number is smaller and more useful because it is the part somebody can go and get. McKinsey puts the value generative AI could unlock for retailers at \$240 billion to \$390 billion, worth something like 1.2 to 1.9 points of margin across the industry. Only a slice of that comes from decisioning systems, about 0.2 to 0.4 points of EBIT margin. That slice changes how the business runs.

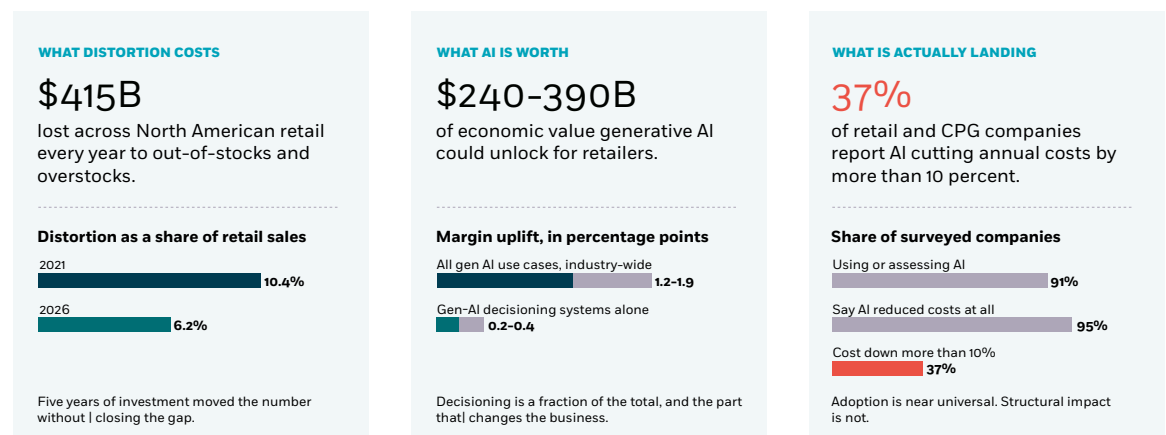
Then there is the awkward part: adoption isn't the constraint. In NVIDIA's 2026 survey of retail and consumer goods companies, 91% were using or assessing AI, and 95% said it had helped bring annual costs down. That sounds like a win right up until you look at the size of it: only 37% saw costs fall by more than 10%. Most of the rest report a rounding error and call it a program. McKinsey's own retail survey, run in April 2024, found something starker still. Ninety percent of executives were experimenting with generative AI. Two out of fifty-two reported successful implementation across the organization.

Two out of fifty-two – not a particularly inspiring figure.

Almost anyone can get a model working now, yet very few can get the organization to run differently because of it. The easy use cases are easy for a reason. You can generate copy, summarize a report, stand up a pilot, and have something impressive to show by quarter end without touching anybody's decision rights. The dirty work sits in the high-frequency decisions that set margin, inventory, supplier terms, labor, and service. Getting at those requires a different opening question. Not which technologies should we deploy, but which decisions in this business create the most work and leak the most money.

The market in numbers

US retail keeps buying intelligence. The gap is in what happens after the insight arrives.



The value is not blocked by a shortage of AI use cases. It is blocked by how slowly the business turns a signal into an action.

Sources: IHL Group inventory distortion research, 2025 and 2026 studies. McKinsey & Company, LLM to ROI: How to Scale Gen AI in Retail. NVIDIA, State of AI in Retail and Consumer Packaged Goods, third annual survey, January 2026.

Figure 1. Three numbers that frame the opportunity: what distortion still costs, what AI is worth, and how little of it is reaching the P&L.



THE PRODUCTIVITY PARADOX

More intelligence has not meant less effort.

We've spent 30 years getting more sophisticated and more complicated at the same time, and it is not obvious the two have canceled out.

Merchants have richer demand signals than ever. Supply chains have visibility that would have read as science fiction in 2005, pricing moves daily, assortments are wider, customer journeys span more channels, and fulfillment networks flex in ways that used to be impossible. To run all of that, retailers added planning tools, reporting environments, analytics teams, control functions, specialist roles, and a great deal of coordination.

Here is the catch. The technology arrived, but the work did not leave. Instead, it compounded.

A report gets automated, and people still meet to talk about it (could this meeting have been an email?). A forecast becomes algorithmic, and the planner still overrides it. A pricing engine makes a recommendation, and an analyst rebuilds the scenario in a spreadsheet before anyone will act on it. Inventory becomes visible, and the supply chain team still fixes the imbalance by hand. The dashboard flags the problem beautifully. Somebody still has to decide what to do, then persuade four other departments to play their part.

Retail feels this more than most industries because it is decision-dense. Price, stock cover, assortment, supplier orders, markdown timing, fulfillment routing: thousands of small calls, repeated across tens of thousands of SKUs, hundreds of stores, and every day of the year. Two minutes of unnecessary effort on any one of them is nothing. At retail scale, two minutes is a department. It's worth being precise about that, because it explains how a design flaw nobody would notice in a single workflow can quietly fund an entire floor of analysts, planners, and coordinators whose real job is to compensate for fragmentation.

The frontier is not better insight. It is less work between the insight and the action.



WHY NOW

We have been here before, and it is worth saying so out loud.

If this argument feels familiar, that is because it should. In 1990, Michael Hammer published “Reengineering Work: Don’t Automate, Obliterate” in Harvard Business Review, and his point was almost exactly the one I am making here: lay technology over a broken process and all you get is the same brokenness, faster. He wanted to obliterate the process and rebuild it around the outcome. The examples still hold up. Ford cut the data it matched in accounts payable from fourteen items to three and removed 75% of the headcount from the function. Mutual Benefit Life collapsed a thirty-step, five-department, nineteen-person application process into a single case manager and cut the timeline from five to twenty-five days to two to five.

Everything since has been a version of the same idea. ERP integrated data in the nineties. Lean and Six Sigma stripped waste out of individual functions through the 2000s. Analytics and demand planning gave everyone better inputs in the 2010s. RPA automated keystrokes in the years after that. Every one of those waves delivered something real. Not one of them removed the sequence.

That is the part worth sitting with for a second. Each wave made a step faster. None of them could make the steps unnecessary, because the sequence existed for a reason. Every function held a different constraint in its head, and the only way to reconcile five constraints was to walk the decision past five people. Hammer’s fifth principle was to link parallel activities instead of integrating their results. He was right in 1990. It was also unbuildable for about thirty-five years. Being right early is not the same as being useful.

What has changed is not that machines can do more tasks. It is that a system can now hold an entire cross-functional decision in context at once: demand, margin, supplier capacity, inventory, customer behavior, and the policy constraints, all of it at the same time, and reason across the whole thing. Earlier waves automated pieces of the workflow. They could not think about the decision. That is the difference, and it is why this is a 2026 argument rather than a rerun of a 1990 one.



WHERE THE COST ACTUALLY COMES FROM

Four layers of cost, and not one of them has a budget line.

Decision latency. Retail decisions routinely move more slowly than the economics beneath them. A competitor cuts price. A promotion runs hot or cold. Inventory starts building up in the wrong region. A supplier flags a constraint. If your response depends on a weekly report, a planning calendar, and three approvals, every day of delay creates work. Pricing investigates the variance, planning reforecasts, supply chain rebalances, procurement revises the order, finance explains the margin miss, and somebody's leadership team reviews it all. Labor is the visible cost. The invisible cost is lost sales, avoidable markdowns, safety stock nobody needed, air freight, supplier leakage, and margin that quietly went somewhere else. You will find the first number in a budget. You will not find the second anywhere.

Process fragmentation. The same decision gets processed several times because each function only sees its own slice. A promotion starts in merchandising, then moves to planning, supply chain, procurement, finance, and marketing. The business still pays six teams to build and reconcile six versions of one answer. Every handoff adds an assumption, a chance to disagree, and a path for the decision to travel backward. Six versions of one answer, and somebody still has to pick.

Manual intervention at scale. Planners override recommendations. Analysts rebuild the same report every Monday. Procurement compares the same suppliers against the same criteria. Finance reconciles data that should never have diverged. Supply chain investigates exceptions that follow a predictable pattern. The problem is that the operating model asks people to repeat the same work at a frequency that ties labor to complexity rather than to value. Add volume, and you add people. That is the whole equation, and it is a miserable one to be locked into.

Organizational drag. Fragmented work eventually coalesces, and that gets sticky. If six functions touch a decision, someone must coordinate them. If two reports disagree, someone must reconcile them. If recommendations need repeated review, a management layer forms to handle the reviews. Give it a few years, and you have reporting teams, PMOs, coordinators, business partners, and managers whose real contribution is holding the seams together. These roles are the consequence of a design decision nobody remembers making. Nobody ever approved it. It accumulated.

The hidden cost stack

Headcount is the visible cost. These four layers are what create it.

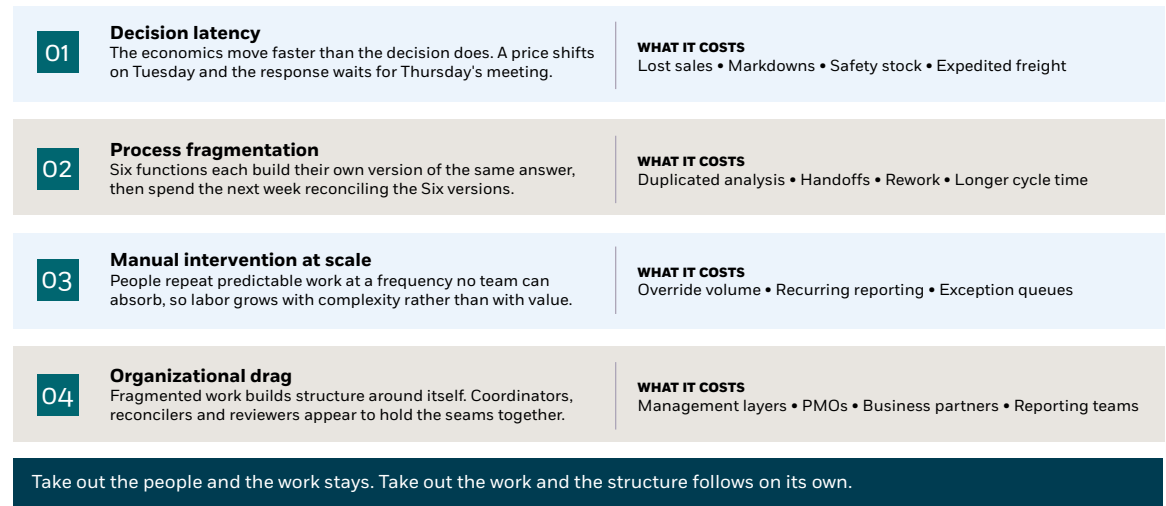


Figure 2. The hidden cost stack. Every layer generates headcount, and none of them ever shows up in a budget as a cost driver.

HOW LATENCY COMPOUNDS

A jacket that missed its plan, and the loop nobody designed

A markdown on a women's jacket sounds like the simplest decision in the building. It is week six of the season. The buyer sees full-price sell-through running at 31% against a plan of 45%, weeks of supply climbing, and proposes taking 25% off in the doors where it is sitting.

Simple, right? Well, let's watch it travel.

Merchandising builds the case from sell-through by door, size, and color brokenness, what the competition is already doing, and twenty years of judgment. The first instinct is usually to give it one more week, because nobody wants to hand back margin on a style that might still turn. Planning then models what the markdown does to the season margin plan and the open-to-buy.

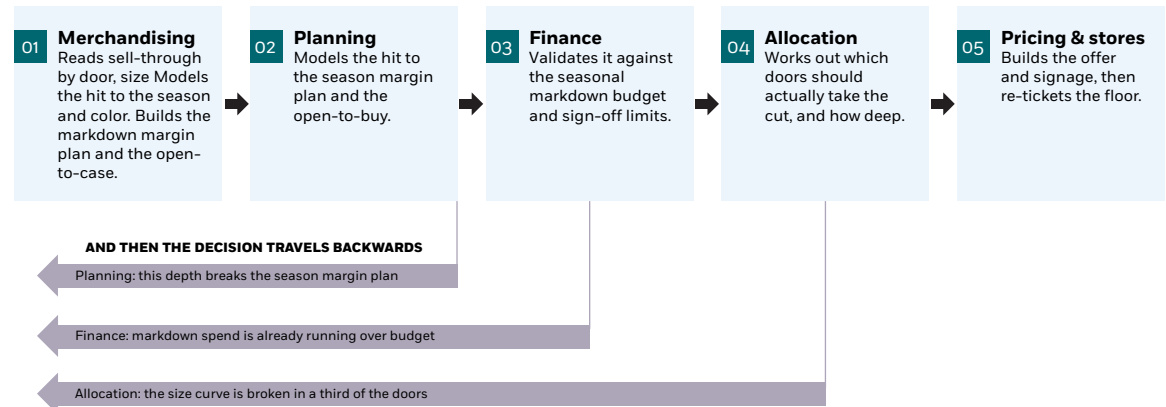
This is usually where the first constraint appears. The proposed depth breaks the margin plan, so the decision travels backward to merchandising for a narrower cut.

Finance does the same thing from a different direction. Cumulative markdown spend is already running ahead of the seasonal budget, so the case goes back again. Then allocation finds the size curve is broken in a third of the doors, which means a chain-wide markdown hands margin away in the stores where the jacket would have sold through on its own. Back it goes once more, this time to be split by door cluster. And, on and on we go.

The expensive part isn't the waiting. It is that every late challenge sends the whole thing 'round again, and each lap adds analysis, meetings, emails, a revised door list, a rebuilt margin case, and another review while the season keeps running underneath it. And nothing in that chain is wrong. Planning should protect the margin plan. Finance should protect the markdown budget. Allocation should stop us from giving margin away in doors that never needed the cut. Those controls are the point. The flaw is purely one of sequencing: the operating model discovers constraints one at a time instead of weighing them together when the decision is made.

One markdown, five functions, and a decision that keeps coming back

A women's jacket at week six, missing plan, traced from the first read to the price a customer actually sees.



Each return trip means fresh analysis, a revised door list, a rebuilt margin case and another review, while the season keeps running and the jacket keeps aging.

The problem is not that five functions touch the decision. It is that the constraints are discovered one at a time instead of considered together.

Illustrative flow based on a standard in-season apparel markdown cycle.

Figure 3. One markdown, five functions, and three separate routes back to the start.

By the time the new price hits the floor, the delay has already cost something. A deeper cut than the jacket needed. Signage and re-ticketing done twice. Weeks of prime floor space held by a style that should have been clearing. And apparel is unforgiving about this in a way replenished categories are not. There is no next week to sell into. The season ends, the floor set changes, and whatever is left competes against newness at whatever price it will move.

What the loop actually costs

It is worth putting numbers on this because the labor everyone can see is the small part.

Walk the jacket through a calendar. Merchandising takes five days to build the case, including the week nobody likes to admit to, where we wait for one more read. Planning takes six to model the margin and open-to-buy impact and to check for brokenness by door. Finance takes four to validate it against the seasonal markdown budget. Allocation takes four to determine which doors should take the cut. Pricing, marketing and stores take five between them to build the offer, produce the signage, update the site and re-ticket the floor. That is twenty-four working days if nothing goes wrong.

Something always goes wrong. Planning sends it back because the depth breaks the margin plan: six days. Finance sends it back because markdown spend is already over budget: five days. Allocation sends it back to be split by door cluster: four days. Call it thirty-nine elapsed days from the sell-through miss to a customer seeing the new price.

Now scale it. At any point in the season, a mid-size apparel retailer has about four hundred styles in the same queue, missing plan and awaiting a decision. Say they average 2,500 units still on hand against a \$48 ticket. That is roughly \$48 million of retail value awaiting a markdown call.

Here is where it costs. Apparel merchants live on the gap between initial markup and maintained margin, and a healthy gap is five to eight points. If it's consistently wider than that, either your promotional cadence or your buy depth is wrong. Our jacket carries a 58% initial markup and a plan that expects to give back six of those points. Act within a week,

and you take 25% off a narrow set of doors and land close to plan, about seven points of erosion. Wait three weeks, and it takes 30% across more doors, and you are twelve points down. Wait five weeks, which is where thirty-nine days puts us, and it takes 40% chain-wide, and you are eighteen points down. Wait seven, and you are cutting 50% into everyone else's clearance, twenty-three points down. Past that, the floor set changes, and the jacket competes with newness at whatever price will move it.

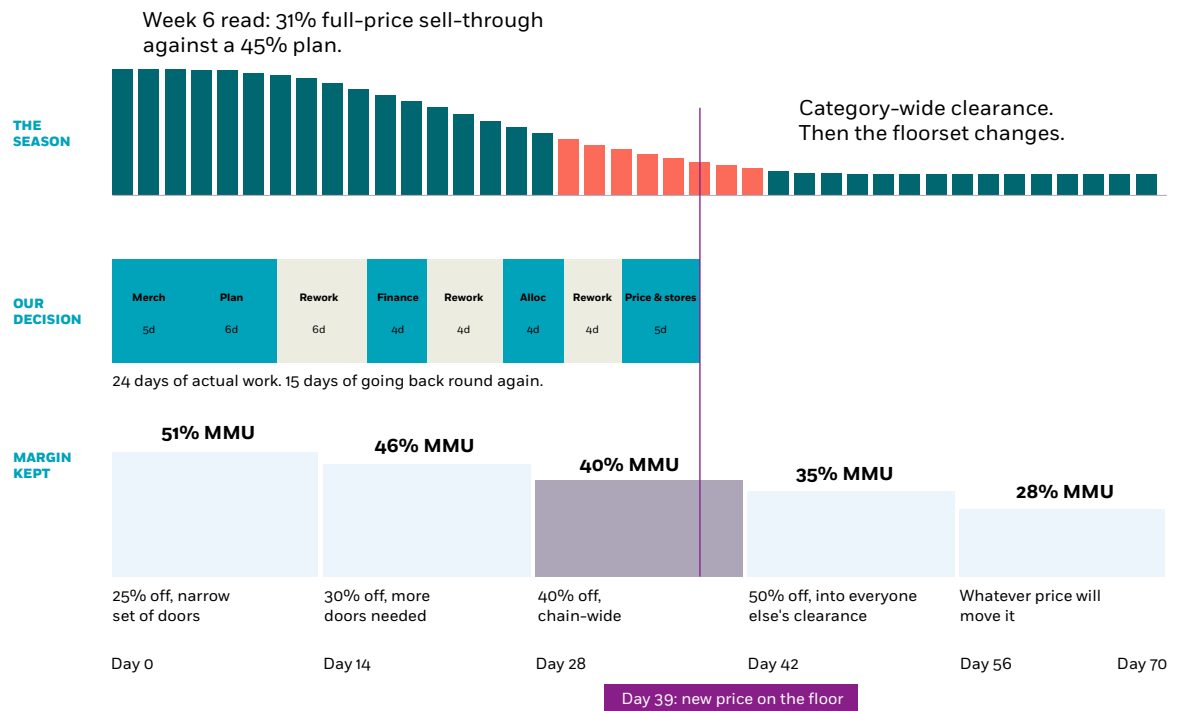
Eighteen points of erosion against a plan of six is twelve points of margin lost to the calendar rather than to the buy. On \$48 million of retail sitting in the queue, that is about \$5.8 million a season. Run it across the six drops most apparel retailers now work to, and decision latency is worth north of \$30 million a year, without anyone making a single bad buying decision.

The per-lap number is the one a CFO should remember. Each trip back adds about five days, and in the steep part of that curve, five days costs a little over two points of margin, or roughly a million dollars across the styles sitting in the queue. The labor inside that lap, five functions, a couple of people each, most of a working day, costs about \$4,500.

So every trip around the loop burns roughly \$4,500 of visible cost and about a million dollars of invisible cost. That ratio is the whole argument of this paper in a single line. Cost programs go after the \$4,500 because it has a name, an owner, and a cost center. The other number never shows up anywhere, because nobody books a journal entry for a decision that arrived late.

The season runs on its own clock. The decision runs on ours.

One jacket, from the week-six sell-through miss to the price a customer finally sees. Directional.



Planned erosion was six points below initial markup. At thirty-nine days it is eighteen. Those twelve points were not spent, they were missed.

Figure 4. The season clock against the decision clock. Directional model: swap in your own initial markup, ticket price, units, and drop cadence, and the arithmetic still lands somewhere uncomfortable.

In a compressed model, the merchant still exercises judgment, and every function still protects what it is there to protect. The difference is that the constraints are considered together rather than discovered one after another. That is how the loop gets broken.



PATTERNS

What separates the programs that hold from the ones that unwind.

What works

Start with the decisions, not the org chart. Headcount benchmarking will tell you that you are heavier than your peers. It will not tell you why, and a benchmark has never fixed anything. Start with the decision, and the answer surfaces on its own: repeated analysis, split ownership, manual exceptions, duplicated controls, and coordination. Once that work is visible, you can tell the difference between the capacity you need and the capacity that exists because the process is badly designed.

Map what people do, not what the procedure says. Documented processes describe intent. A useful baseline captures manual touches, handoffs, approvals, queue time, rework, exception rates, spreadsheet activity, meetings, and how often the same information gets rebuilt from scratch. Process mining helps when the system data is good. Sitting with the team helps everywhere else, because most friction lives in email, spreadsheets, and the corridor. Find the spreadsheet somebody maintains at home on a Sunday night. That is where your process lives.

Redesign the flow, not the function. Finance can speed up its validation step while the promotion process stays just as slow. Supply chain can improve forecast accuracy while commercial keeps making late changes. Everybody hits their number, and the calendar does not move an inch. The gains come from redesigning the whole path around the outcome, which means asking where the decision starts, what it needs, who owns it, which constraints belong inside it, and where a human genuinely has to intervene.

Put the guardrails in early. Margin thresholds, inventory limits, supplier capacity, service rules, and risk tolerance should increasingly live inside the decision itself. That doesn't remove accountability; it moves it forward. Now, these functions are doing it through agreed logic that prevents a predictable problem, instead of another round of validation after the fact. The control still exists. It just stopped costing you three weeks.

Keep people where judgment earns its keep. This isn't a race to remove humans. Business judgment, supplier relationships, customer impact, regulatory risk, and genuinely unusual exceptions all justify a person. Machines take the volume; people take the cases where context, negotiation, ethics, creativity, or real money is involved.

What does not

Taking out heads while leaving the workload behind. This books a saving in year one and creates a problem in year two. The remaining team inherits the same reports, approvals, exceptions, and coordination load. Decisions slow further, service slips, managers absorb the overflow, and eighteen months later the roles return under different titles. I have watched this happen more than once. Savings hold only when you remove the work or properly simplify it before removing the role.

Automating a step without questioning the process. A faster reconciliation is useful. It is also just a faster step in a fragmented workflow. If the same functions still touch it, in the same order, with the same controls, you have improved a task and left the cost-to-serve problem exactly where you found it. Congratulations, you are now late faster.

Adding dashboards without moving authority. Very few retailers lack information. A dashboard that identifies a pricing issue is worth nothing if nobody has the authority to act on it, if the output still must be argued across three functions, or if the decision waits for Thursday. Insight without decision rights is another reporting layer, and you already have plenty of those.

Letting every function keep its own version of the truth. Different perspectives are valuable. Conflicting economic facts are not. When commercial, finance, planning, and supply chain each maintain their own demand, margin, inventory, and supplier assumptions, reconciliation becomes a permanent feature of the operating model. At that point, you are not debating strategy. You are debating whose spreadsheet is right. Agreeing on shared facts does not mean one team wins. It means the trade-offs are made explicitly rather than through a chain of handoffs.

Treat adoption as the outcome. Users, prompts, models, pilots, and automated tasks are activity metrics. They can look a lot like momentum even when the economics sit still. Ask what moved on the P&L and watch the room go quiet. AI becomes strategically relevant when cycle time, cost, margin, inventory, service, working capital, or human effort changes.

Very few retailers are short of information.
Most are short of the authority to act on it.

THE CONVERSION PROBLEM

A retailer can automate a third of a function's workload and save nothing.

This headline is intentionally blunt because I cannot overstate how often this happens. If the structure doesn't change afterward, automation adds capacity without reducing costs. That's not to say capacity isn't worth having. It can improve service, absorb growth, strengthen analysis, or handle the next volume increase without a hiring spree. It just shouldn't be counted as a savings number until something has moved on the P&L. A disciplined cost-to-serve program tracks four separate events and refuses to blur them:

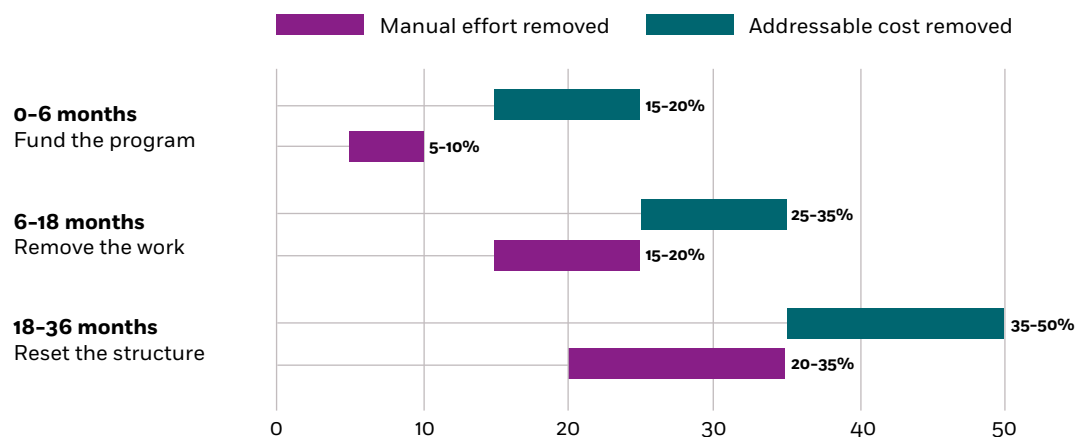
- **Work identified.** The activity exists and has been measured.
- **Work eliminated.** The activity no longer happens, or happens far less often.
- **Capacity released.** Named hours are freed.
- **Value captured.** The financial consequence is visible in the accounts.

In merchandising, that might mean fewer manual forecast overrides and fewer promotional scenarios. In finance, less recurring reporting and reconciliation, and only then a change to the reporting organization. Supply chain and procurement follow the same order, with routine controls, expediting, supplier analysis, and exception handling reduced before anyone resets role requirements.

The last step is a management choice and should be made out loud. Freed-up capacity can come out of the cost base, move to higher-value work, or be held to support growth without future hiring. All three are defensible, but they are not the same benefit and should never be counted as such. Finance should keep a value ledger that separates hard cost removed, capacity released, cost avoided, margin improvement, and working capital, so the program cannot accidentally take credit three times for the same win.

Effort comes out first. Cost follows later, and only on purpose.

Automating work creates capacity. Capacity becomes savings only when someone decides what to do with it.



The blue bar always leads the amber one. That gap is capacity waiting for a decision.

Track four separate things: work identified, work eliminated, capacity released, value captured. Only the fourth one reaches the P&L.

Illustrative ranges. Actual conversion depends on starting maturity, outsourcing arrangements, labor mix, technology estate and workforce decisions.

Figure 5. Effort reduction consistently runs ahead of cost reduction. The gap between the two bars is capacity created but not yet converted.

EVIDENCE

Nobody has built the whole model. Several have proved the pieces.

No retailer has solved this end-to-end, and any vendor claiming otherwise is selling something, probably snake oil. What good operators are demonstrating is narrower and more useful: individual parts of the model deliver real economics as soon as the technology changes the workflow around them.

Walmart's Self-Healing Inventory is a good example of preventing work rather than processing it. The system spots an imbalance and redirects stock to the stores that need it. Walmart reports the capability has generated more than \$55 million in savings. The mechanism is more instructive than the number. The organization no longer has to discover, analyze, escalate, and manually coordinate a category of problem that was always predictable.

Macy's makes the workflow point. RFID gave the company item-level accuracy, but the return came from rebuilding the store routine around it. An independent study found stores using RFID to locate and sell merchandise running about 6.1% ahead of control stores, and merchandise sitting in the back room instead of on display falling from 30% to somewhere between 4% and 6%. The tags were necessary. The new floor routine paid off. Worth noting that Macy's did this more than a decade ago, which is rather the point. The mechanism is not new. The willingness to redesign around it still is.

Amazon sits at the other extreme of automation and lands in the same place, with up to 25% faster processing at its next-generation fulfillment sites and a 10% improvement in robotic travel efficiency from DeepFleet. Robotics, AI, inventory systems, and workflow redesign were treated as one operating system, not four programs with four steering committees.

Three companies, three starting points, one pattern. In every case, the technology was the cheap part. The economics came from a willingness to change who does what and when.

Nobody has built the whole model. Plenty have proved the pieces.

Three US operators, three different starting points, one consistent pattern in the economics.



In every case the technology changed the work itself. None of the savings came from putting a smarter layer on top of the old routine.

Sources: Walmart, U.S. Supply Chain Playbook Goes Global (2025). Platt Retail Institute study of the Macy's RFID deployment. Amazon, next-generation fulfillment and DeepFleet operations updates.



WHAT WE ARE DOING ABOUT IT

This is the work we do at UST, and it is not a tools business.

I'll spare you the sales pitch, but I'd be doing you a disservice if I laid all this out and didn't say where UST fits in.

As the strategy and management consulting arm of UST, Evolve helps retailers become AI-ready in the sense this paper means it: not more models, fewer steps between a signal and an action. Most of what we do in an engagement isn't the model at all. It's mapping the decision as it happens, finding the handoffs and approvals that exist only because something upstream was fragmented, redesigning the workflow around the outcome, and then doing the unglamorous change management that decides whether anybody uses the new way. AI accelerators are easy to buy and easy to strand. Ours come wrapped in process and workflow redesign, because that's what determines whether the released capacity ever shows up on the P&L.

Two of our products make the point better than I can in the abstract.

UST Concert goes after the content supply chain. A large brand manages 50,000-plus content assets across retailers, formats, and markets, and the traditional way of producing them runs on photoshoots, agencies, and manual approval chains. Concert generates high-fidelity lifestyle imagery, video, and copy at the source, turning weeks into minutes. But the number that matters isn't render time. It's the routing. The brief, the shoot, the agency round, the brand review, the legal review, the retailer-specific rework: that's the bloat, and most of it exists because assets used to be expensive to make and therefore expensive to get wrong. Make them cheap to produce and cheap to vary, and the approval chain built around scarcity has nothing left to protect. Thousands of retailer-specific variants stop being a resourcing conversation.

UST Concept-to-Store does the same for the planning calendar, the machine that produced our jacket. Traditional development is sequential: forecast, commit, produce, then find out. Concept-to-Store uses AI-generated mood boards and designs to decouple design from manufacturing, so concepts are generated and tested in parallel instead of queued. We scope these engagements around a 30 to 50% reduction in concept-to-launch timelines, smaller initial production runs, and less capital tied up in bets nobody validated. Look at what that does to Figure 4. Product hits the market closer to peak demand, full-price sell-through goes up, and markdown exposure comes down. Same twelve margin points, recovered at the front of the process instead of defended at the back.

Neither of these is a tool we install and leave. Both change who does what and when. That's the only test this paper accepts.

FUNDING

Design the economics before the technology roadmap.

Many large transformations fail on sequencing rather than substance. Technology gets funded first, then integration happens, and finally, if it ever happens at all, operating redesign.

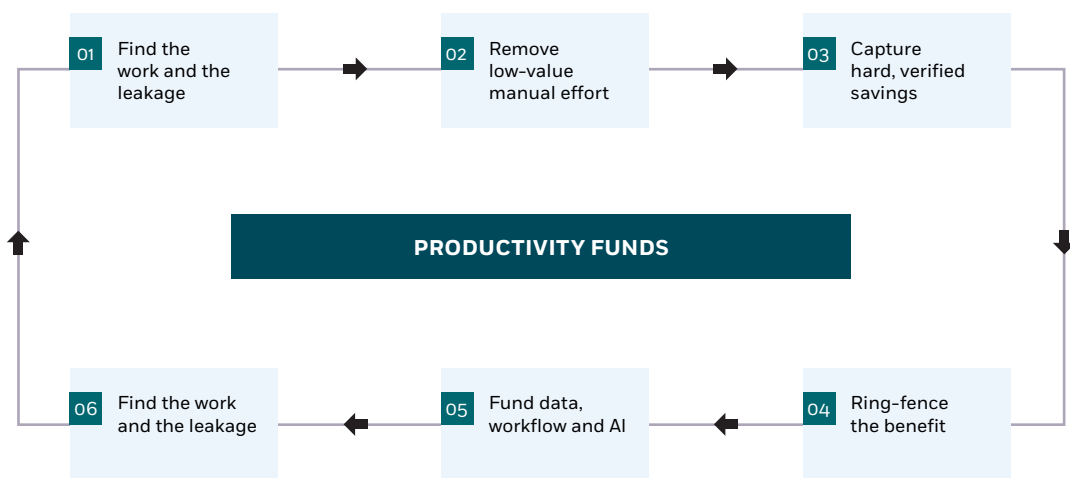
The better model uses the current cost base to help pay for the future one. Early waves target work that can be simplified without waiting for a platform rebuild: recurring reporting, reconciliations, unnecessary controls, and weekly exceptions. None of that needs a platform. It needs somebody with the authority to say stop. Those savings are captured and ring-fenced rather than swallowed by general budgets, and they fund the data, integration, workflow, and AI capability needed to remove the bigger cross-functional processes in the next wave.

This is as much a finance discipline as a technology one. A transformation office should set phase gates for reinvestment and hold the program to the value ledger described earlier. Every new investment should have a clear line to the work it removes and the value pool it opens. If you cannot draw that line on a napkin, you are not funding a transformation. You are funding a hope.

The flywheel gets stronger as the program moves upstream. Early simplification releases pockets of capacity. Better decisioning removes more of the workflow. Cross-functional redesign takes out duplication. Only then does structural change capture savings that weren't available at the start, because the work underneath still existed. Each turn buys the next one, and that is the only version of this I have seen survive a budget cycle.

A transformation that pays for its own next phase

Early savings are ring-fenced and spent on the capability that removes the next, larger pool of work.



The discipline here is financial as much as technological. If the saving is not ring-fenced, it quietly disappears into the general budget and the second wave never gets funded.

Figure 7. The self-funding flywheel. Each turn buys the capability required for the next one.



THE OTHER HALF OF THE ARGUMENT

Compressing decisions doesn't just cut costs. It goes after revenue the old model couldn't reach.

Everything up to this point is a cost argument, and I think it's a good one. It's also only half the story, and, honestly, the equally interesting half.

Go back to the jacket for a second. Those twelve margin points didn't vanish because somebody was lazy or bad at their job. They vanished because the season kept moving while the decision sat still. That's not a cost line anywhere in the P&L. That's margin we simply never had the chance to keep. Every day of latency is another day of pricing, assortment, and availability calls being made against a market that already changed underneath us. Compress the decision, and you don't just spend less deciding. You land inside the window where the decision was still worth making.

Then there's the volume problem, and this one gets less airtime than it deserves. The old operating model could only make decisions at whatever grain a human chain could physically chew through. In practice, that meant the category, the region, the week. Not the store, the SKU, the customer, the hour. Nobody sat down and chose that. It was a throughput ceiling dressed up as a strategy. Take the ceiling away, and a whole class of decisions becomes economically possible for the first time: localized assortment, store-level pricing, substitution logic that reflects what this shopper will accept instead of what somebody assumed about the segment six months ago. That's revenue the old model couldn't have produced at any headcount. You couldn't hire your way to it.

The third piece is the one that changes the clock. The customer is starting to show up as software. McKinsey puts it at three to five trillion dollars of consumer commerce mediated by AI agents by 2030, and their warning to retailers is refreshingly blunt: if your catalog, policies, and value proposition aren't machine-readable, agents and therefore shoppers simply won't find you. It doesn't matter how much people love the brand.

So think about what that does to a three-week decision cycle. An agent doesn't respond to a campaign. It reads your price, availability, delivery promise, substitution rules, and returns policy, stacks them against everybody else's, and decides. Then it does the whole thing again an hour later. In that market, every operational decision this paper treats as a cost driver stops being overhead and becomes the product. Price reliability, fulfillment predictability, a substitution policy that makes sense: that's what gets selected. Brand storytelling matters less than operational trust, because the thing doing the shopping can't be charmed.

Which is where decision compression stops being an efficiency program and turns into a commercial capability. The retailer who can reprice, reallocate, and re-promise inside a day takes business from the retailer who needs three weeks, and takes it without either of them running a campaign. Nobody sends a press release about it. The cost case is what pays for the work. The demand case is why we can't sit on it.



GETTING STARTED

Pick one decision. Not a catalogue of use cases.

We're not short on AI ideas. I've got a dozen; you've got a dozen. Many retailers have a list longer than a CVS receipt. The hard part is deciding which two or three deserve real attention, engineering capacity, and where they'll exhaust their political capital.

I'd advise picking the transformation at the domain level first, before picking use cases, and then backing it with cross-functional teams, modular architecture, real data quality, and genuine upskilling. Spread the same money and the same engineers across thirty disconnected pilots, and the arithmetic never works, which is roughly what the two-out-of-fifty-two is telling us.

A sensible starting point is one decision-rich area where the economics are meaningful and the work already crosses several functions. Promotion planning, markdowns, replenishment, supplier planning, returns, assortment, and the commercial period close all qualify, because each one combines measurable value with visible organizational friction. Rewire a single important flow deeply enough to prove the model, rather than spreading scarce data, engineering, and change capacity across thirty small experiments.

- 1. Map the decision as it happens.** Document the trigger, required data, people involved, systems touched, handoffs, approvals, queue time, exception rates, rework loops, and the economic cost of delay. For a promotion, follow it from the first commercial idea through demand planning, supplier capacity, margin validation, activation, and post-event cleanup. The output is not another process map for the shelf. It is an understanding of where decisions stall, where information is rebuilt, and which activities exist only because something upstream is fragmented. That becomes the baseline against which you measure everything else.
- 2. Agree on the value case before choosing the technology.** Decide what financial or operational outcome this decision is supposed to improve. It might sit in margin, sales, inventory, working capital, labor productivity, waste, supplier terms, service, or some mix of those. Finance should validate the current baseline and define how benefits will be recognized, because hours saved is evidence of productivity and nothing more. A credible case uses the same five-way split set out earlier, with a named owner against every line.
- 3. Choose the level of autonomy the way a manager chooses how much rope to give.** Not every decision should jump straight from human execution to full automation. A high-risk markdown, a material supplier commitment, or a decision that lands unevenly across customers may warrant human review, even when the underlying analysis is fully automatable. Management already solved this problem, and the solution has a name.

The Six Levels of Delegation is a widely taught framework for how a leader delegates work, and it runs like this. Level one, do exactly as I say. Level two, investigate this and report back. Level three, give me your recommendation, and I will decide. Level four, decide, but check with me before you act. Level five, decide and act within these limits. Level six, get it done. What makes it useful is what it is not. It is not a grade for how capable somebody is in general. It states what this person has demonstrated, on this kind of decision, to this leader. Our staircase is that same ladder pointed at a decision system rather than a person:

Assist (levels one and two):

The system gathers, prepares, and reports. A person decides and acts.

Recommend (level three):

The system proposes an action. The person accepts, adjusts, or rejects it.

Act with approval (level four):

The system prepares the execution and waits for a signature before anything moves.

Act within guardrails (level five):

The system executes routine cases inside agreed limits and escalates the rest.

Autonomous (level six):

The system executes and adapts within boundaries proven over time.

Autonomy should rise with evidence, not with enthusiasm

The Six Levels of Delegation, applied to a decision system. Each rung is earned on one decision type, for one accountable owner.

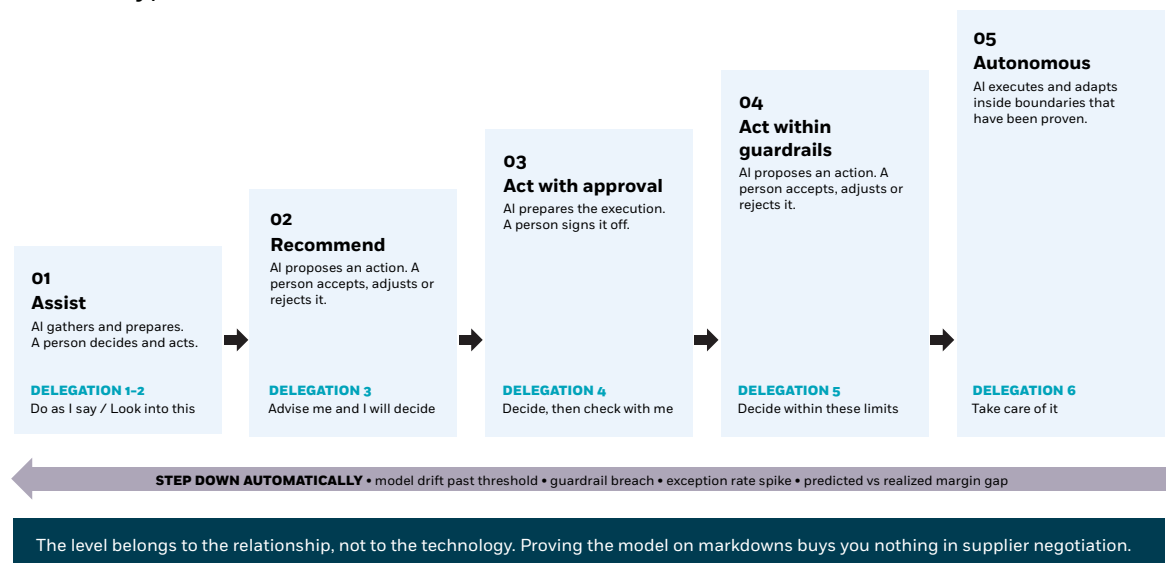


Figure 8. The autonomy staircase mapped to the Six Levels of Delegation, with the step down most delegation frameworks never draw. Risk, financial materiality, reversibility, and customer impact decide how far up a given decision belongs.

Where the analogy earns its keep, and where it breaks

Five rungs against six levels, and the gap is where it gets interesting. The mapping isn't one-to-one, and where it fails is instructive. Levels one and two are genuinely different jobs. Doing exactly this is execution. Looking into it is investigation, and what separates them is initiative. For a machine, they collapse into the same rung because retrieval and preparation cost it nothing and require no initiative. Machines start higher on the ladder for gathering and lower on judgment than people. Worth remembering the next time somebody tells you an AI is as good as a junior analyst. It is better than one at part of that job on day one, and it will never be good at other parts of it.

The most valuable rung is the one nobody buys. At level two, look into this for me. It barely exists in most AI deployments. Retailers jump from nothing straight to level three because a recommendation demos well and a watchlist does not. Go back to the jacket. That thirty-nine-day clock started at the week-six review because that is when a human happened to look. A system at level two, doing nothing more ambitious than watching four thousand styles and raising a hand when a sell-through curve bends, starts the same clock at week three. It does not decide at all. It just removes the most expensive part of the timeline: the part before anybody knew there was a decision to make.

With a person, the level only goes up. With a system, it must go both ways. Promoting somebody is a good day. Demoting them is a significant management event that most leaders put off far longer than they should, which is why delegation frameworks are almost always drawn as a one-way staircase. A decision system has no feelings, and that is an advantage worth designing for. Write the step-down triggers before go-live: model drift past a threshold, a guardrail breach, a spiking exception rate, or a sustained gap between predicted and realized margin. Then let the system drop a rung on its own and tell somebody it did. A staircase people can watch you walk down is one they will let you walk up.

The level belongs to the relationship, not to the technology. This is what confuses executives most. The same model, deployed in two business units, settles at different autonomy levels, and everyone goes hunting for a technical explanation. There usually is none. In the delegation literature, the level is a property of a particular leader, a particular person, and a particular task, which is why two managers sit at different rungs with the same employee. It works the same way here. Autonomy is not a feature you procure. It is an organizational state you earn, one decision type and one accountable owner at a time, and it does not transfer. Proving the model on markdowns in womenswear buys you nothing in supplier negotiation.

Where it breaks: a person can explain themselves afterward. Every analogy has a limit, and this one has an obvious one. You can ask a person why they made a particular call last Tuesday and get an answer a merchant, an auditor, or a regulator will accept. You cannot do that with a model after the fact. So the delegation framework never had to carry one requirement: the evidence must be captured at the moment the decision is made, not reconstructed later. What data was in front of it, which alternatives it weighed, which guardrail applied, and whose name is on the outcome. With a person, you can ask afterward. With a system, if you didn't log it, it didn't happen.

4. **Redesign the workflow at the same time.** Dropping AI into an existing process while leaving every approval, role, meeting, and control intact is the most reliable way to disappoint. Treat the new capability as the reason to remove steps, move controls earlier, change decision rights, simplify policy, and reset service expectations. If a model can weigh demand, margin, supplier capacity, and inventory simultaneously, there is no reason to preserve four sequential validation meetings that existed only because a person could not.
5. **Decide in advance how capacity becomes value.** Design the workforce consequences before deployment, not after. If an AI-enabled workflow frees up a fifth of a planning team's capacity, leadership must decide whether to release that capacity, redirect it to higher-value work, or absorb it into future growth. HR, finance, and the business owner should align early on role changes, timing, skills, and organizational implications. Otherwise, you land in the most common failure mode: a genuine productivity improvement, no change in the cost base, and no change in what people spend their days doing.



A MISSING DISCIPLINE

Forcing AI into everything is an excellent way to blow your budget.

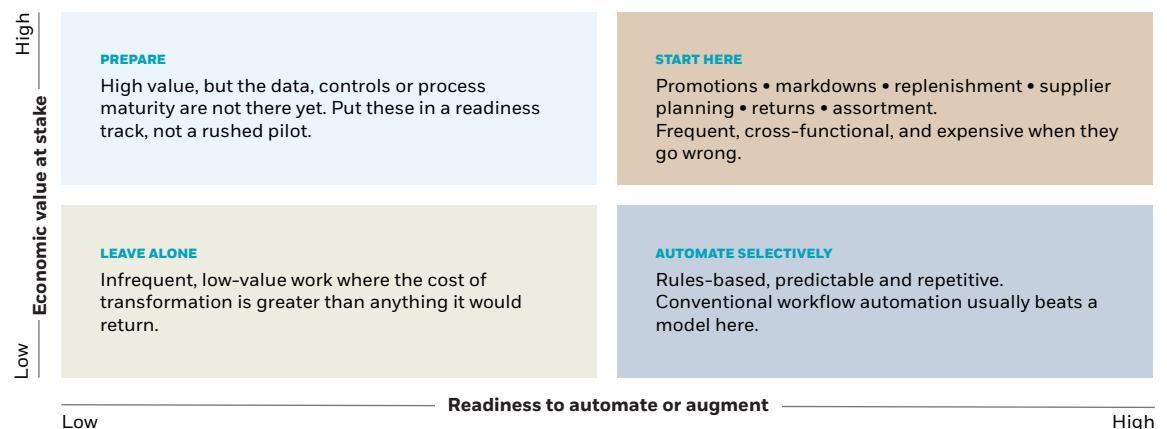
One of the fastest ways to burn an AI budget is to treat every inefficient process as an AI opportunity. Some work should stop. Some processes need a policy change, not a model. Deterministic, rules-based work is usually better served by conventional workflow automation or by functionality already sitting unused in a platform you have paid for. Look. It is usually in there. AI earns its place when a decision carries enough volume, variability, data, and economic consequence to justify something more intelligent.

The strongest candidates are the ones that happen often. They require someone to weigh several signals at once and often cross functional lines. Delay or inconsistency costs real money, and the business can define what an acceptable action looks like. Pricing, promotions, markdowns, replenishment, assortment, supplier planning, returns, and inventory exceptions usually fit. Infrequent decisions with poor data, unclear ownership, highly bespoke judgment, or thin economics are usually not the best choice and are better left human-led until the surrounding process matures.

That gives you a portfolio discipline rather than a wish list. High value and high readiness go first. High value with weak data or controls goes into a readiness track rather than a rushed pilot. Low value but highly repeatable work gets automated selectively with simpler technology. Low value and low readiness gets left alone, which is a decision, not a failure. Saying no to a use case is not a lack of ambition. It is what keeps the other three from starving.

Start with a decision, not a catalogue of use cases

The best first move combines real economics with enough data, control and ownership to actually scale.



Not every inefficient process is an AI problem. Some work should simply stop, and some of it just needs a rule.



OWNERSHIP AND GOVERNANCE

The business owns the outcome. Technology enables it. Governance keeps it scalable.

Ownership is one of the most common reasons AI programs stall, and the failure modes are predictable. Park the initiative entirely in IT, and the business never changes the workflow. Park it inside a single function, and you get a patchwork of models, vendors, and data flows that nobody can govern. Park it in an AI center of excellence with no P&L accountability, and you get an impressive portfolio of pilots with no economic owner. I have seen all three. The third one always has the best slides.

The model that works separates three things: outcome ownership, technical ownership, and enterprise governance.

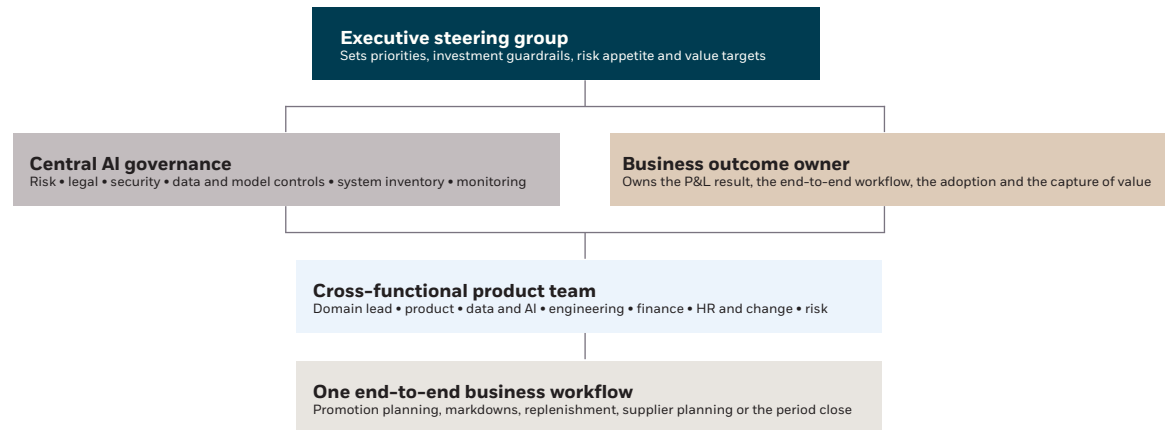
A senior business leader owns the economic result and the end-to-end workflow. That person needs real authority to change decision rights, KPIs, policies, roles, and handoffs across every function involved. Real authority, not a dotted line and a monthly forum. For promotion planning, that usually means a commercial executive owns the outcome, even though supply chain, procurement, finance, marketing, and technology all shape the design.

A cross-functional product team owns delivery and continuous improvement, combining domain expertise, product management, data and AI, engineering, finance, HR or change, and risk. Technology and data leaders own the platform, integration, architecture, security, and reusable capabilities. Finance validates the value case and benefit realization. HR owns workforce transition and capability building. Central governance sets the risk, model, data, monitoring, and third-party standards. Write those names down. If you cannot assign a person to each one, you do not yet have an operating model.

The executive steering group should stay out of individual decisions. Its job is portfolio governance: priorities, capital, risk appetite, enterprise dependencies, and whether the benefits are landing. This distinction is not pedantic. Governance that gets too operational recreates precisely the decision latency the program was set up to remove. You would be rebuilding the exact thing you are paying to remove, except now it has a charter and a standing invite.

Who owns it, and who keeps it safe

Business ownership sits closest to the economics. Governance stays central enough to be trusted everywhere.



Centralize the guardrails. Decentralize the ownership. Keep the steering group out of individual decisions.

Figure 10. Separating outcome ownership, delivery and enterprise governance.

Governance without a new bottleneck

The model must be risk-tiered. Low-risk, reversible decisions can operate with broader machine authority once performance is proven. Decisions with meaningful customer, legal, financial, or reputational consequences keep tighter human-in-the-loop control. The aim is pre-approved boundaries, so routine activity moves fast and unusual or material cases reach the right people quickly. Decide the boundaries once, in daylight, rather than case by case in a Thursday meeting.

This all comes down to a manageable set of controls: an inventory of material AI systems and agents, documented business owners and decision rights, data provenance and access control, testing and validation before release, performance and drift monitoring after deployment, human appeal and override, incident response, and a defined process for retiring or replacing a model. These belong inside product delivery, not bolted on at the end as a compliance exercise. Bolt them on at the end, and you will spend the last six weeks of the program doing archaeology.

Centralize the guardrails. Decentralize the ownership.



FOUNDATIONS

AI cannot compress a decision if the business still has to reconcile the data first.

Every technology conversation should start with one question. Can the workflow reach a trusted version of the information it needs to act? A promotion can't be priced by a model that can't see supplier funding, and in most retailers, that funding sits in a system no one has ever connected to anything. If product, inventory, supplier, customer, price, promotion, and financial data live in disconnected systems, the AI layer inherits the fragmentation beneath it, and you end up with a sophisticated model whose output still must be checked by hand before anyone believes it. At that point, you have bought a very expensive suggestion. Four things need to hold.

Shared data products. Nobody needs to complete a perfect enterprise data transformation before starting, and if you wait for one, you will be reading this paper again in five years. But the decision being redesigned does need reliable inputs. Work backward from that decision and build governed, reusable data products for the information that matters most. Make the quality expectations explicit enough that the workflow itself knows when the data is trustworthy and when to raise an exception.

Workflow and orchestration. Insight with no route to action has no economic value. The architecture needs a way to connect models to operational systems through APIs, events, workflow engines, and, where it fits, agentic orchestration. What matters is the path itself: signal, then recommendation, then approval or execution. If the last mile is a person retyping something into another system, you have not finished.

Observability and control. The retailer needs to know what the model decided, what data it used, whether performance is drifting, and how a person intervenes. Logging, monitoring, override, auditability, threshold management, and incident response are product features, not governance paperwork. Build them in the sprint, not in the audit.

Modular architecture. The technology will keep moving quickly, so avoid depending on one model or one vendor when you can. Separating business workflow, data products, orchestration, models, and user experience lets you replace a component without rebuilding the process around it. Whatever is state of the art the day you sign will look quaint by the time you finish.



THE LAST MILE

Getting the technology working is about halfway there.

Retail transformation consistently underestimates how much organizational change is required after the build is finished, and the reason is usually in the incentive structure. Not in the technology. In the comp plan.

A planner still measured by how many forecasts they personally review has no reason to trust an exception-based model. A finance team still accountable for rebuilding every commercial business case will continue validating decisions by hand, even when the economics are already embedded. They are doing exactly what you are paying them to do.

Role design must evolve alongside workflow design. Merchants should spend more time on category strategy, customer relevance, assortment, and supplier economics. Planners should shift from touching every forecast to managing material exceptions and feeding business context back into the model. Procurement should shift capacity from administrative buying to negotiation, strategic sourcing, and supplier innovation. Finance should move closer to the decision, embedding guardrails upfront and concentrating human effort on challenges, capital allocation, and the genuinely unusual. If the job description does not change, neither does the job.

The workforce plan also needs a realistic transition path. Some capacity will be released, some will be redeployed, and some roles will require genuinely new skills. Attrition, spans and layers, outsourcing arrangements, and organizational timing all shape how quickly productivity gains translate into cost reductions. HR should be in the room from the beginning, with a view of future roles, skills, training, redeployment, and where structural reductions are expected. Bring them in at the end, and you will get the transformation you deserve.

Management routines matter just as much as job descriptions. Weekly meetings meant to review what the system already monitors should be rebuilt around exceptions and decisions. KPIs should measure outcomes, not activity. Leaders must be explicit that the new workflow is how the work gets done, or teams will run the old process alongside it. The capacity everyone was promised will never appear—two processes, twice the work, and a very confusing dashboard.

The future workforce is not defined by fewer people. It is defined by far less human effort spent on coordination and correction.

A PRACTICAL PATH TO SCALE

Fund it, remove the work, then reset the structure.

Horizon	Objective	Typical focus	Directional effort impact	Directional cost opportunity
0–6 months	Fund the program	Planning and merchandising, finance, supply chain, procurement	15–25% reduction in targeted manual effort	5–10% reduction in targeted addressable cost pools
6–18 months	Remove the work	Promotions and pricing, forecasting and replenishment, supplier planning, cross-functional finance	25–35% reduction in targeted cross-functional effort	15–25% cumulative opportunity
18–36 months	Reset the structure	Exception-based operating model, management layers, planning and control roles	35–50% reduction in routine effort in highly automated processes	20–35% cumulative opportunity

All ranges above are directional. They are planning hypotheses meant to frame the size of the prize, not benchmarks. Rebuild them against your own process baseline, cost base, technology estate, outsourcing model, labor economics, and risk appetite before anyone puts them in a business case.

The first six months establish the decision and cost baseline, simplify the obvious, and capture benefits that don't depend on a large systems program. Planning overrides and Monday reporting packs are usually the two fastest wins, followed by the validation loops nobody defends and the exceptions that recur every week. The point is to build enough credibility and cushion to fund what follows. Nobody funds year two on a promise.

Between six and eighteen months, the program moves upstream into the cross-functional decisions that create the work. Pricing and promotion mature, forecasting becomes more continuous, replenishment connects more directly to demand, procurement receives better forward signals, and finance guardrails move inside the decision. This is where whole pieces of workflow come out, rather than individual tasks getting faster. It is also the phase everybody wants to skip.

From eighteen to thirty-six months, the structure can reset to reflect the smaller amount of work that remains. Management layers can flatten because there is less to coordinate. Obviously, the exact savings will vary enormously by retailer, but this sequence should not.



THE SCORECARD

Four measures, tracked together, that are hard to game.

Decision productivity. End-to-end cycle time, number of human touches, handoffs, approval steps, rework loops, exception volume, and the share of decisions completed within agreed service levels. These metrics show whether the workflow is getting simpler or just becoming more digital. The difference matters, and only one pays—economic outcomes. Hard cost removed, capacity released, avoided hiring, margin improvement, inventory and working capital, waste, supplier economics, cost-to-serve and service. Finance should own the benefit methodology so that the program cannot confuse real P&L impact with productivity that hasn't been converted yet. Let the P&L owners keep score.

Adoption and operating model change. The proportion of work running through the new workflow, the rate and the reasons for manual overrides, role and span changes, user confidence, and whether legacy processes have been switched off, not quietly kept alive by three people who never trusted it. A new model running beside the old process produces no structural productivity at all.

AI performance and control. Model quality, decision accuracy, drift, guardrail breaches, human overrides, incident rates, and how each autonomy level is performing. The objective is to concentrate human intervention where it improves the outcome or reduces material risk, not to drive it toward zero. Zero would worry me.



THE END STATE

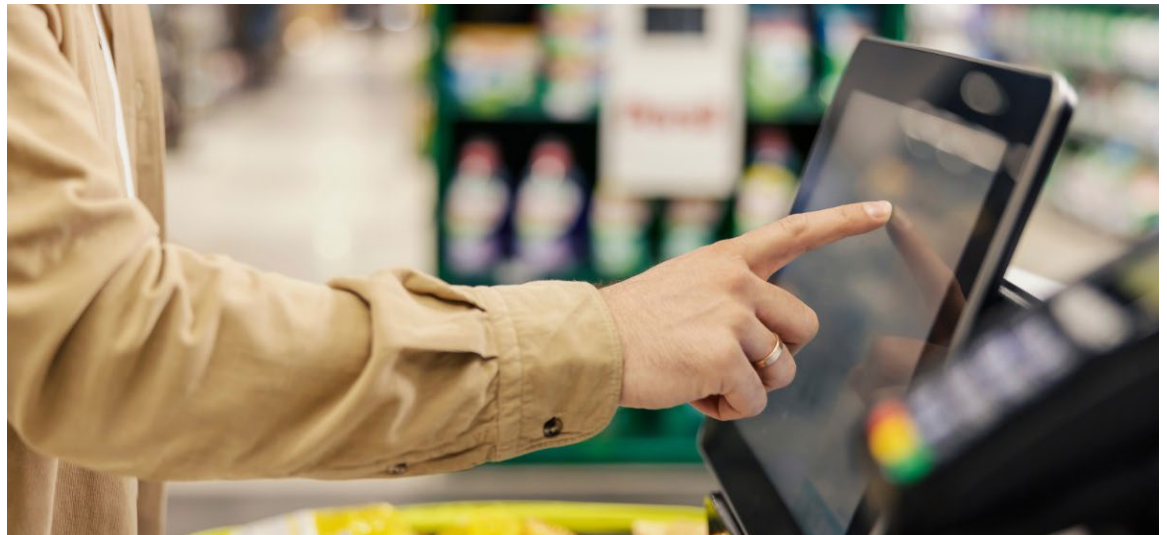
The winner is not the retailer with the most AI. It is the one that needs the least effort to act.

The strongest retailers will still have merchants, planners, buyers, finance leaders, supply chain experts, and managers. What changes is how those people spend their week. Far less time will go to gathering information, reconciling versions of the truth, preparing routine analysis, moving decisions between departments, and fixing predictable problems after they happen. Nobody got into merchandising to reconcile a spreadsheet.

The clearest way to see it is in where the hours go. Today, a large share of merchant, planner, and finance time is spent producing and reconciling the inputs to a decision rather than making it. In a compressed model, that share collapses, and the same people spend their week on the part of the job they were hired for in the first place—same people. Better week.

This is also why the autonomous retailer is the wrong ambition to hold up. Full autonomy is neither necessary nor desirable for every decision. The more useful goal is a low-friction operating model where routine volume moves quickly within trusted guardrails, exceptions reach the right people early, and human judgment is concentrated where it genuinely adds value. Ambition is fine. Just point it at the right thing.

Get there, and the support organization no longer needs to grow at the same rate. Better decisions generate less downstream correction. Productivity becomes a recurring source of investment rather than an occasional cost exercise. And the cost base becomes more resilient because the work that used to support it has gone. That is the version of this that survives the next downturn.



FINAL THOUGHTS

Modernization and productivity are not the same thing.

The biggest risk for retailers is not moving too slowly on AI. It is layering AI onto an operating model that never changes. You can have better forecasting, smarter pricing, automated reporting, richer analytics, and a more advanced supply chain, and still need the same meetings, handoffs, approvals, reconciliations, and management layers to reach a decision. That is modernization without productivity, and it is expensive.

The retailers that pull ahead will systematically reduce the effort required to turn information into action. They will use AI where intelligence genuinely changes the decision, simpler automation where the work is predictable, and human judgment where context or risk warrants it. They will also decide in advance what happens to released capacity: where it will be reinvested and where cost will come out. Put that decision in writing before anyone becomes emotionally attached to the answer.

None of this requires a big bang. The credible path is narrow and sequential. Start with a small number of decision-rich workflows, capture hard value early, and use the savings to fund stronger data and orchestration. Then move progressively upstream into the decisions that create the work. Boring, sequential, and effective.

For years, retail transformation has been judged by what was implemented. The better test is simpler and much harder to fake. Three years from now, can the business handle the same volume of decisions with fewer people in the room, and can finance point to the line where that showed up?

If you've read all the way to the end, you might have one last question—"In the beginning, you said you got the first of two things you wanted. What was the second?" The second came a few days later—a signed engagement letter.

Modernization photographs beautifully in a board deck. Productivity shows up in the numbers.

Contact us

**Gregory Serrago**

Head of Retail Transformation | UST Evolve

gregory.serrago@ust.com

**Eric Pilkington**

Group Chief Executive and General Manager | UST Evolve

eric.pilkington@ust.com

Global HQ: 101 Enterprise, Suite 300, Aliso Viejo, CA 92656 | **Phone:** +1 949 716 8757

Since 1999, UST has worked side by side with the world's best companies to make a powerful impact through transformation. Powered by technology, inspired by people, and led by our purpose, we partner with our clients from design to operation. Our digital solutions, proprietary platforms, engineering, R&D, products, and innovation ecosystem turn core challenges into impactful, disruptive solutions. With deep industry knowledge and a future-ready mindset, we infuse expertise, innovation, and agility into our clients' organizations—delivering measurable value and positive lasting change for them, their customers, and communities around the world. Together, with 30,000+ employees in 30+ countries, we build for boundless impact—touching billions of lives in the process.

ust.com