The Growth Consulting Playbook

How Firms Grow When Software Delivery No Longer Pays the Bills

Lee Harrington

Contents

The Growth Consulting Playbook

How Firms Grow When Software Delivery No Longer Pays the Bills

Lee Harrington

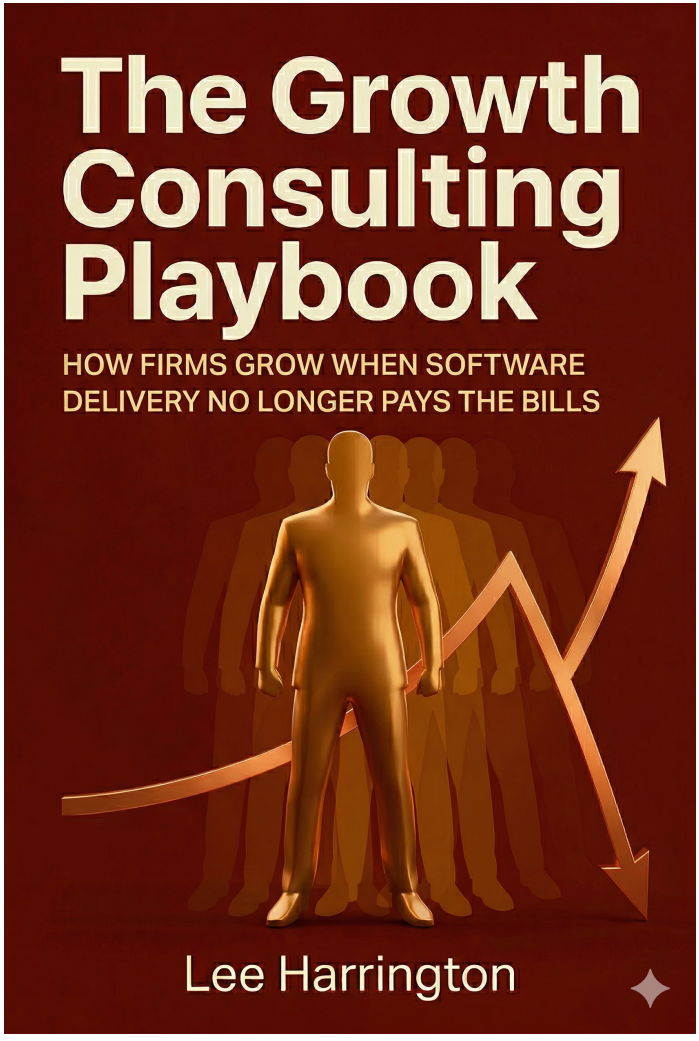

Figure 1: Cover

Introduction — The Wrong Win

The first AI win is easy to recognize. A team that used to need three weeks now needs five days. A proposal that used to eat a weekend gets done in an afternoon. A delivery manager looks at the new velocity, looks at the payroll, and sees a clean line from productivity to cost reduction.

Of course they do.

If you lead a consulting firm, or even a practice inside one, this is the first temptation you have to fight. The numbers seem to line up. The story is easy to tell. AI made us more efficient. We took out cost. We improved margins. We got ahead of the market.

Sometimes that story is even true for a quarter or two. The problem is what happens next, because what disappears is rarely visible on the org chart.

The spreadsheet can show you salary cost. It cannot show you the person who knows why a client keeps objecting at the last minute. It cannot show you the delivery lead who can smell a bad requirement before it becomes a bad build. It cannot show you the architect who can translate between the sales promise, the technical constraint, and the political reality inside the client organization. It definitely cannot show you the future capability you just cut before it had a chance to matter.

That is the wrong win.

This book is about what happens when consulting firms confuse the first visible AI gain with the real strategic opportunity. For years, a large part of consulting economics rested on labor. Not just expertise. Labor. Hours. Headcount. Leverage. The mix of senior people who sell and shape the work, mid-level people who manage and translate it, and junior people who execute enough of it to make the math work.

AI does not destroy all of that overnight. It does something more specific and

more dangerous. It puts pressure first on the most bounded, repeatable, and
legible parts of delivery: routine research, standard implementation work, boilerplate code, predictable analysis, the parts of the job that used to justify large amounts of billable effort because they took real time and real hands. That is enough to change the business model.

Not because all software delivery stops mattering. It does not. Not because every developer or consultant suddenly becomes replaceable. They do not. And
not because consulting is dead. It is not.

The real shift is subtler and more important. Value starts moving. It moves away
from undifferentiated delivery labor and toward judgment, orchestration, context, trust, redesign, and the ability to turn AI-enabled productivity into something larger than cost savings. It moves toward the people and firms that
can help clients become more capable, not just leaner.

That is the argument of this book. But before we get to the opportunity, we need to name the trap clearly.

The Efficiency Trap

The trap is simple enough that smart people fall into it all the time. AI makes
delivery cheaper, clients want savings, consultants know how to sell savings, and everyone gets rewarded for acting quickly.

That is why the early phase of the AI era is so deceptive. It does not begin with a clean upward curve where firms get more productive and naturally turn
that productivity into stronger businesses. It begins with a dip. Old delivery economics weaken. Firms feel margin pressure. Leaders rush to show that they
are responding. Some of them cut. Some of them cut too deeply. Some of them
call that discipline.

What they are doing is eating the seed corn.

That phrase matters because it describes the mistake better than most management language ever will. If you eat the seed corn, you solve the immediate problem and weaken the next harvest. You convert future capacity into
present relief. The move can still feel rational. It can even look responsible. That is why it keeps happening.

The same thing happens in consulting when AI gains are harvested only as labor
reduction. You improve short-term economics. You also cut continuity, tacit knowledge, client context, informal governance, and the apprenticeship path that creates your next generation of high-value operators.

You can save yourself into a smaller future. You can also help your clients do
the same thing and then wonder why the market for your old work keeps getting
thinner.

The AI Dip and Divergence Curve

This book treats the early AI era as a real model, not just a mood.

The first shape is a dip. Confusion rises. Delivery gets repriced. False wins appear. Some firms look faster before they become stronger. Others look more
efficient before they become more brittle. There is a contraction phase here, and pretending otherwise will make smart leaders late to the real decision.

Then the paths diverge.

On one path, firms take the reduction-first route. They use AI to remove labor,
capture shallow savings, and protect this year's margin. They may still do well
for a while. But they also weaken continuity, shrink their developmental pipeline, and make it harder to create the kind of differentiated value clients will still pay for when generic delivery gets cheaper everywhere.

On the other path, firms take the multiplier route. They still pursue productivity. They still care about efficiency. But they do not treat efficiency as the destination. They protect the people who carry strategic context. They redesign work. They redirect capacity into better offers, deeper
client work, faster innovation, stronger internal systems, and new forms of growth.

Both paths may begin with the same tool. They do not end in the same firm.
That divergence is where this book lives.

What This Book Actually Is

This is not a prompt book. It is not a survey of AI vendors. It is not a generic future-of-work book. It is not a celebration of automation, and it is

not a complaint about automation either.

It is a strategy book for consulting leaders, especially founders, managing partners, practice leaders, and business builders in boutique-to-mid-market firms who can feel the delivery economy changing under their feet and need a
better response than "do more with less."

If you are in a large global firm, this book is still for you. The economics are different at your scale, but the pressure is not imaginary. If you are an independent consultant, it is also for you. You may not have a partner group or
formal service lines, but you still face the same strategic choice: keep selling undifferentiated labor, or move your value up a level. In some ways, the choice is even sharper when the firm is just you, because there is nowhere to hide from commodity pressure and no one else to redefine the offer for you.

One point is worth stating plainly. I have not run a global consulting firm, and I am not writing this book as a managing partner defending a résumé. I am
writing it from more than three decades spent inside consulting systems across
global firms, boutiques, and client-side transformations, as an architect, practice builder, presales leader, delivery lead, commercialization lead, and builder of centers of excellence. I have spent enough time close to the work to
see what these firms reward, what they miss, and what they cut too easily. That is a different vantage point. For this argument, I think it is an honest one.

The reason this book focuses on consulting is not that consulting is the only place this shift matters. It is that consulting is one of the first places where the old model breaks in a way smart people can no longer ignore. Consulting sits right at the fault line between knowledge work, client trust, billable labor, and business-model design. It is where the old leverage system meets the new reality fastest. That makes it a good place to see the future clearly.

What Gets Cheaper, What Gets Scarce

Once you stop looking at AI as just a speed tool, deeper questions show up. What kind of work is actually becoming cheap? What kind of work is becoming more
valuable? Which people become easier to replace, and which become more important than before? Where does the productivity go?

If the answer to that last question is only "into cost savings," you are using new capability to make the future smaller.

This is where many firms have not adjusted their mental model. They are treating AI as a better engine for the same vehicle: faster research, faster code, faster analysis, faster slide production. All of that matters. But speed alone does not tell you what business you are in now.

When routine delivery compresses, the scarce thing changes. Judgment matters more. Context matters more. The ability to redesign work matters more. The ability to preserve and transfer knowledge matters more. The ability to help clients grow, not just slim down, matters more.

The firms that understand this early will not simply use AI better. They will be selling something different.

What You Cannot Afford to Cut

The strongest argument against reduction-only strategy is not that cuts feel cold. It is that they are strategically clumsy.

When leaders remove people, they do not only remove labor cost. They remove memory, trust, pattern recognition, relationship continuity, informal teaching, and the human excess capacity from which new capability often emerges.

That is why one of the central ideas in this book is **Knowledge Carriers**. These are the people whose value is larger than their visible output. They hold context the organization has never fully written down. They connect functions that do not naturally talk to one another. They know how a client actually works, not just how the statement of work says the client works. They can tell the difference between a problem that looks technical and a problem that is really political, relational, or structural.

A firm that cuts those people because AI improved a dashboard or wrote a first draft faster is making the exact mistake it thinks AI will save it from: mistaking legible output for actual value.

This book will argue that AI does not just force firms to ask who can be cut. It forces them to ask who they did not realize they were depending on, and

who
they did not realize could become far more valuable once given leverage. Most
firms never ask that question. That is the gap this book is written to close.

What This Book Promises

The ideas in this book have to survive Monday morning. They have to help you
see the trap sooner, protect the right people, and decide where the new
capacity should actually go.

By the end of the book, you should be able to do five things with much more
clarity than most firms have right now:

1. recognize the Efficiency Trap before you build strategy around it
2. see why the first AI win can be the wrong win
3. know what not to cut when routine delivery gets cheaper
4. decide where freed capacity should go instead of letting it disappear into
   shallow savings
5. redesign your firm's offers, engagements, and client conversations around
   growth rather than reduction

They have to be usable in a partner meeting, in a pricing
conversation, in a headcount review, in a service-line strategy session, and in
the quiet moment when a consulting leader realizes the market has already
changed faster than the firm has.

There is an economic argument in this book, because there has to be. But there
is also a human one. The firms that navigate this transition well will not do
it by pretending people are a sentimental concern floating around the real
business problem. They will do it by recognizing that people are where more of
the real business value now lives.

That is the work ahead.

The next chapter starts where the pressure starts: with the end of the delivery
economy as consulting firms have known it.

Chapter 1 — The Delivery Economy Is Ending

The consulting business did not grow for decades by selling wisdom alone. It grew by wrapping wisdom around labor.

That is not a cynical statement. It is just a clear one. A firm won work because clients trusted its judgment, brand, and experience. But the economic engine under that trust was still built on people doing the work. Teams were assembled. Hours were billed. Junior and mid-level consultants learned by producing drafts, building models, writing code, documenting systems, testing flows, cleaning data, and carrying implementation forward one task at a time.

The model worked because the labor was expensive to assemble, expensive to coordinate, and expensive to replace. Even when clients complained about rates, they still lived in a world where real delivery required real human throughput. There was no shortcut around that fact.

That world is changing.

The shift is uneven. Some service lines will feel it sooner. Some firms will hide from it longer. But enough of the old economics are weakening that consulting leaders need to stop treating this as a normal cycle.

This Is Not a Downturn Story

Consulting firms know how to read a soft market. They know what happens when clients freeze budgets, delay projects, or move cautiously after a bad quarter. They know what it feels like when demand dips but the basic model still holds. You tighten hiring. You protect the pipeline. You wait for spending to come back.

That is not the whole story here.

A cyclical slowdown says, “the work is still valuable, but demand is softer for now.” A structural shift says, “the market is changing what kind of work it is willing to pay premium prices for.”

That is the difference that matters.

When AI helps teams produce first drafts faster, generate boilerplate code more
cheaply, automate routine research, compress analysis time, and reduce the manual burden of implementation, the client does not just see a better tool. The client sees a different price anchor for delivery itself.

That matters even when the AI output is imperfect. In some cases it matters especially then. The market does not need AI to be magical for delivery economics to change. It only needs enough visible compression in time and effort for clients to start asking a harder question: why are we still paying for this the old way?

Once that question takes hold, repricing follows.

The timing still matters, and this is where many firms confuse a structural shift with instant maturity. AI is more like a horse than a car. A horse has opinions, moods, and a learning curve on both sides of the relationship. A good rider on a good horse is a sight to behold. Most other combinations are a
disappointment. That is how new technologies usually behave in real organizations too. They get worse before they get better because the tools are
still improving, the humans are still learning, and the operating model around
them is still immature.

Consulting leaders have seen this before. Plenty of firms adopted agile on troubled programs because they wanted to move faster. The first project was
not better. Often the second one was not either. The learning curve was real.
The same thing is happening here. The question is not whether AI will change
consulting economics. It will. The question is whether your firm is developing riders or just buying horses.

What Is Actually Being Repriced

Sloppy claims make smart readers stop listening, so this needs to be precise.

This book is not arguing that all consulting work is suddenly cheap. It is not
arguing that all software delivery becomes commodity work. It is not arguing that clients no longer need expert builders, architects, translators, or operators.

It is arguing something narrower and more consequential: routine, bounded, repeatable, legible delivery work is under the greatest pressure first, and a

large enough share of consulting revenue still depends on that work that the business model cannot shrug it off.

That includes things like:

- standard implementation tasks
- predictable analysis
- boilerplate or pattern-heavy code
- documentation and first-draft production
- test scaffolding
- workflow mapping that follows familiar patterns
- junior-level research and synthesis that used to consume days

None of those activities disappear. But many of them become faster, easier to
scope, easier to benchmark, and harder to defend at yesterday's price.

You can hear the repricing in a client conversation before you see it in a market report. A buyer looks at a proposal and says, "I understand why this
used to take six people. I do not understand why it still does." Nothing in that sentence says they think the work is trivial. It says they no longer accept the old relationship between effort and price.

A delivery economy starts ending when that sentence starts showing up in more
rooms.

The Problem with the Old Pyramid

Traditional consulting leverage depended on a shape. A relatively small number
of senior people sold and framed the work. A broader middle translated, managed, and quality-controlled it. A larger base carried a meaningful share of
the production load.

That pyramid was not only an org chart. It was a training system and a pricing
system at the same time.

Junior work mattered because it was billable, because it taught the next layer
of consultants how the work actually got done, and because clients accepted that certain kinds of progress took time. The base of the pyramid was not just
cheap labor. It was economic fuel and developmental infrastructure.

AI puts pressure on that base first.

If the early-draft work, first-pass analysis, standard build work, and other repeatable tasks require fewer people or fewer hours, the traditional leverage model starts to wobble. Revenue pressure appears before the firm has rebuilt its offers. Apprenticeship weakens before leadership has replaced it with a new development path. The old engine keeps running for a while, but it is no longer running on the same fuel.

This is why the chapter is not called "Delivery Is Changing." It is called "The Delivery Economy Is Ending." The shift is not only technical. It is economic.

Why Clients Will Push This Harder Than Firms Expect

Clients do not have to become AI experts to put pressure on consulting margins. They only have to notice that some work now seems faster and more available than it used to be.

They will notice.

Some already have. Internal teams are using AI to create first drafts that once required outside help. Technical leaders are seeing demos of work compressed from weeks to days. Procurement teams are hearing every provider claim efficiency gains. Boards and CFOs are being told AI should improve productivity. Once that expectation enters the system, it does not stay politely confined to one function.

It moves into pricing conversations. It moves into staffing assumptions. It moves into scope debates. It moves into the silent comparison every buyer makes between what they think a team should cost and what a proposal still says it costs.

Consulting firms are being surprised by a change they can already describe. They understand the technology intellectually, but many still talk about their offers as if the market will keep valuing visible effort the way it used to.

It will not.

Delivery Still Matters. Delivery Alone Does Not.

This is the point where some readers will want to object, and it is worth meeting that objection honestly.

Complex programs still fail because requirements are muddy, organizations are
political, systems are messy, and implementation is hard. Clients still need teams who can build, integrate, govern, adapt, and recover when reality refuses
to match the slide deck. None of that disappears because an AI tool generates a
first draft quickly.

Good. Keep that objection. It is partly right.

It is worth lingering there for a moment, because this is where some of the best firms will misread their own strength. They know their work is hard. They
know messy programs still need experienced people. They know clients still pay
for recovery, trust, judgment, and execution under pressure. All true.

What changes is not the need for serious delivery. What changes is the share of
the fee that can still be justified by visible effort alone. The harder the real work becomes, the less sensible it is to describe the value as hours, hands, and throughput. That language undersells what the client is actually buying.

So the objection does not rescue the old economics. It points to the new ones.

If what remains scarce is not raw production but judgment under constraint, coordination across messy realities, trust, design, prioritization, and the ability to turn productivity into meaningful business change, then that is where value is migrating. The winning firms will still deliver. They just will not be paid mainly for looking labor-intensive while they do it.

Delivery still matters. Delivery alone does not.

That is a hard sentence for some firms because it lands on more than pricing. It
lands on identity. Many firms still think of themselves as premium builders who
happen to advise. The market is forcing a more uncomfortable question: are you
really selling differentiated judgment, or have you been relying on expensive execution to carry more of the business than you realized?

The End of One Engine Creates Pressure Everywhere Else

Once delivery-heavy work starts to compress, the pressure does not stay neatly
inside the delivery function.

It hits hiring because the old utilization assumptions get weaker. It hits training because fewer junior people are needed to produce the first layers of
work. It hits pricing because clients expect some of the productivity gain to show up in the deal. It hits positioning because firms can no longer describe themselves mainly by what they build. It hits sales because buyers start asking
for outcomes, acceleration, and leverage rather than bodies and hours.

It also hits strategy, which is why this book exists.

The wrong response is to stare at the shrinking part of the model and try to protect it with better language. The right response is to recognize that the firm now has to move its value up a level. That does not mean abandoning delivery. It means refusing to confuse delivery with the full story of value.

Some firms will make that move early. Others will keep trying to defend the old
economics one statement of work at a time. They will tell themselves they are
still selling expertise when the market increasingly experiences them as staffed effort with nicer branding. Some of them will also overclaim, talking as if the horse already rides itself because they bought access to a powerful tool. It does not.

That tension gets us to the next chapter.

Because once leaders feel this pressure, the first response is almost always the same: take the easy win, harvest the obvious savings, and call the move strategy.

That is where the trap tightens.

Chapter 2 — The Easy Win

Once leaders feel the pressure described in the last chapter, they almost never
start with philosophy.

They start with math.

That is what makes the next move so understandable and so dangerous.

A team gets faster. Delivery takes fewer hours. Drafts arrive sooner. A few people using AI seem to produce what it used to take a larger team to produce.
Someone does the obvious calculation. If the same output now requires fewer
hands, why would we not reduce the hands?

That question is not foolish. In many rooms, it is the responsible question.

If you are a CEO, a CFO, a practice leader, or a delivery executive staring at
margin pressure, you are not paid to ignore visible productivity gains. You are
paid to do something with them. If a client expects AI savings, if your board
expects AI savings, if your competitors are already talking about AI savings, then reducing labor cost looks less like aggression and more like competence.

That is why the first AI win so often becomes a cost story.

It is visible. It is measurable. It is easy to explain upward. It produces a slide that everyone in the room understands.

This process now takes five days instead of fifteen.
This team now needs six people instead of ten.
This proposal can be delivered with fewer billable hours.
This margin can be protected without raising price.

Every line is legible. Every line feels disciplined.

That is why so many smart firms walk straight into the trap.

Why Cuts Come First

The first reason is simple: savings are easier to prove than growth.

If you cut labor, the number shows up quickly. If you preserve headcount and
try to redirect productivity into future offers, stronger capability, deeper client work, or new revenue, the payoff arrives later and with more ambiguity.

Short-term proof beats long-term possibility in most operating systems.

The second reason is organizational. Cost reduction has an owner. It belongs to
finance, operations, and delivery leadership in ways that feel familiar.
Growth reinvestment is messier. It crosses sales, capability, product design,

talent, service-line strategy, and client development. It demands coordination.
Cuts can be decided in a smaller room.

The third reason is psychological. When leaders feel uncertainty, they reach for the kind of action that signals control. Headcount reduction, utilization tightening, contractor cuts, and scope compression all create that feeling. They say: we are responding. We are being disciplined. We are not asleep.

Those moves can be emotionally reassuring long before they are strategically sound.

The Win Looks Real Because Part of It Is Real

One of the worst ways to argue against the easy win is to pretend it contains no truth.

Of course AI can reduce labor cost. Of course some work can now be done with
fewer people. Of course some bloated delivery models deserve to be challenged.
Of course there are firms carrying work patterns that should have been rethought years ago.

The easy win works because part of it is real.

That matters, because readers will reject this chapter if it sounds like a plea for denial. This is not that. Some cuts will be justified. Some work will not survive repricing. Some roles will narrow. Some forms of staff augmentation will get weaker because they should.

The problem is not that leaders see savings. The problem is that savings become
the whole story before anyone asks what else is being reduced at the same time.

The Shared Dip Rewards the Wrong Instincts

This is where the AI Dip and Divergence Curve matters.

The early phase of a structural shift is not elegant. It is full of noise.
Margins tighten. Delivery assumptions wobble. Sales teams feel pressure. Buyers
push harder. Boards ask what the AI plan is. Leaders do not experience this as
a calm moment for portfolio reinvention. They experience it as exposure.

Exposure creates urgency. Urgency rewards visible action.

The shared dip encourages reduction-first behavior even in firms that would rather believe they are taking the long view. Leaders do not need to be cynical to make short-term moves. They only need to be under pressure inside a system that rewards near-term proof.

This is also why the first wave of AI decision-making can produce so many false wins. A firm cuts quickly, protects this year's numbers, and gets praised for discipline. A client automates a chunk of work, reduces staff, and tells a good story about modernization. The spreadsheet improves before the second-order costs arrive.

By the time those costs become visible, the original decision has already been framed as success.

The Consultant's Incentive Problem

Consultants do not just observe this logic. They often help package it.

This is not because consultants are uniquely shortsighted. It is because the easy win fits the machinery of consulting unusually well. It is easy to scope. It is easy to model. It is easy to sell to an executive sponsor who needs a clear ROI story. It turns AI from an abstract possibility into an actionable program with measurable savings.

In other words, it looks like a good engagement.

If a consulting firm tells a client, "we can help you automate this workflow, remove this layer of manual effort, and reduce the labor required to run it," everyone involved knows how to talk about that work. The buyer understands the case. Procurement understands the math. The sponsor can defend the investment. The consulting team can build a project around it.

That familiarity is exactly the problem.

The more fluently a firm can sell labor substitution, the easier it becomes to help a client shrink the very capability base that would have supported deeper growth later. The consultant gets rewarded for the first win even if the larger system gets weaker.

One reason the trap is so dangerous is that it does not feel like failure while you are doing it. It feels like good consulting.

Eating the Seed Corn Looks Like Discipline

That phrase matters to me because it captures what normal management language
usually hides.

If you eat the seed corn, you do not look reckless. You look practical. You solved the immediate problem. You improved the quarter. You acted while others
hesitated. You can explain every choice you made.

That is what makes the mistake so hard to catch in real time.

Reduction-first leaders are rarely cartoon villains hacking away at a company for sport. More often they are intelligent operators making individually reasonable decisions inside an incomplete frame. The real failure is not intelligence. It is using a frame that counts savings and ignores what those savings consume.

If the only question on the table is "how much labor can AI remove," then the
answer will reliably point toward cuts. If the real question is "what future capacity are we willing to consume in exchange for a short-term gain," the conversation changes.

Most firms never ask the second question early enough.

Why Growth Is Harder to Defend

Growth uses a different grammar from cuts.

Cuts sound exact. Growth sounds aspirational unless it is translated into specific choices. Firms slide toward reduction even when leaders
say they believe in reinvestment.

"We reduced cost by 18 percent" is clean.

"We kept the team, redesigned the workflow, used the freed capacity to deepen
two client relationships, prototyped a new advisory offer, and positioned the firm for stronger revenue next year" may be the better strategy, but it does not fit on a slide as neatly.

It also asks for more courage.

Keeping capacity when the market is telling you to harvest savings feels riskier

than taking the visible win. Redirecting productivity into growth requires a thesis about where the value will come from next. Many firms do not yet have
that thesis, so they fall back to the only logic they can explain cleanly.

This book will keep returning to one practical question:

Where does the productivity go?

If leaders do not answer that question explicitly, the answer defaults to reduction.

Why the Easy Win Becomes the Wrong Win

The easy win becomes the wrong win when it narrows the future faster than it
improves the present.

That is the move many leaders miss. They think they are taking cost out of a
system. Sometimes they are also taking context out, apprenticeship out, continuity out, trust out, option value out, and the human slack from which new
capability often emerges.

They do not see that loss immediately because those things are not measured as
cleanly as labor savings. But the absence shows up later in slower adaptation, weaker relationships, thinner benches, brittle delivery, and a firm that is more efficient on paper than it is capable in practice.

The wrong win is still a win for a while. That is what gives it cover.

But if the first AI gains are harvested without a growth thesis, without a view
of what must be protected, and without a plan for where the new capacity goes,
the firm is not redesigning itself for the future. It is liquidating pieces of the future to make the present look cleaner.

That is the trap.

The next chapter is about what gets lost when that trap closes: not in theory,
but in people, continuity, and judgment that most firms do not realize they are
cutting until it is already gone.

Chapter 3 — What Gets Lost When You Cut Too Deep

If the last chapter explained why cuts come first, this chapter explains why the
story those cuts tell is incomplete.

The easiest thing to measure is labor cost. The hardest thing to measure is what that labor was quietly holding together.

That is why firms get this wrong.

A spreadsheet can show salary, utilization, contractor spend, and margin. It can show which roles look duplicative after automation. It can show who appears
to produce less visible output than the strongest individual contributors on the
team.

It cannot show you what breaks three months later.

It cannot show you the client relationship that becomes harder to steady when
pressure rises. It cannot show you the project manager who knows which stakeholder always objects late and why. It cannot show you the architect who
remembers the compromise everyone made eighteen months ago and can still
explain why it was made. It cannot show you the quiet translator who keeps sales, delivery, security, and the client from misreading one another at the worst possible moment.

Those things are real. They are just not equally legible.

The Organization Knows More Than the Org Chart Can Show

One of the most expensive management mistakes is to assume the formal structure
of the organization tells you where its value lives.

It does not.

Some value is visible. Some is documented. Some is encoded in process, ticket
history, architecture diagrams, CRM notes, and statements of work. But a great
deal of operational continuity lives elsewhere. It lives in memory, judgment, relationships, workarounds, habits of escalation, pattern recognition, and

the
human ability to notice when something small is about to become costly.

That is the layer firms put at risk when they cut too deeply.

This is not sentimentality. It is operational reality.

Organizations run on more than formal authority and visible output. They also
run on informal coherence. Someone knows how the client actually makes decisions. Someone knows which internal dependency is always late. Someone
knows that a requirement written as technical scope is really a political compromise nobody wants to reopen. Someone knows which person can calm a bad
meeting down before it becomes a bad account.

You do not see the value of that knowledge most clearly while it is present. You see it when it is gone.

Knowledge Carriers

This is why the phrase Knowledge Carriers matters.

Knowledge Carriers are the people whose value is larger than their visible output. They carry context the organization has never fully written down. They
hold continuity across projects, clients, systems, teams, and decisions that were never documented well enough to survive clean transfer. They are often the
people who know what kind of mistake the organization is about to repeat because they were there the first time.

Some are senior. Some are not.

That matters.

Firms often assume the most important people to retain after AI will be the most visibly strategic or the most visibly productive. Sometimes that is true.
Sometimes the more dangerous loss is the person in the middle who can translate
between commercial, technical, and organizational reality. Sometimes it is the
delivery lead who can smell trouble before the dashboard moves. Sometimes it is
the person whose name does not come up first in a talent review because their
best work looks like fewer fires, fewer surprises, fewer misunderstandings,

and
fewer reasons for executives to panic.

Those contributions are easy to undervalue because they often appear as the
absence of failure.

A Person Can Look Replaceable Right Before You Need Exactly Them

This is not a theoretical problem.

Years ago, I watched a technically strong consultant drift toward termination because the visible story around him had become simple. The client was unhappy.
The relationship felt rough. The easiest conclusion was that he was not working
out.

That was the visible story. It was not the real one.

The deeper issue was a communication breakdown wrapped around technical work the
client still needed and largely respected. Once that was clear, the job was not
to replace the person. It was to coach the relationship, repair the exchange, and let him reestablish trust directly. That happened. A month later the client
was praising him.

Why does that story matter here?

Because the value at risk was never just task completion. It was continuity, recoverable trust, technical judgment, and a person who could still become more
valuable once the real problem was identified. The original story made him look
like a labor problem. The deeper reality was that he was carrying value the surface read had missed.

That is what firms do to themselves when they cut by visible output alone.

What Gets Removed Along with Cost

When leaders take labor out of a system, they do not remove one thing. They
remove a bundle.

They remove context: the remembered details that let people act quickly without having to rediscover the situation from scratch. They remove continuity: the through-line between old decisions, present constraints, and future moves. They remove pattern recognition: the human ability to say, "we have seen this kind of problem before, and it gets expensive if we miss it early."

They also remove relationship memory, informal teaching, and exception handling. They remove the workarounds that should probably be redesigned but still keep the system functioning until redesign happens. And they remove the human slack from which new capability often emerges, because organizations do not invent much when every remaining person is operating at the edge of confusion.

This is one reason aggressive cuts can make an organization look cleaner before they make it weaker. The cost leaves immediately. The hidden functions linger just long enough for the decision to be declared a success.

Then the second-order effects begin.

Projects require more explanation. Decisions take longer. New people repeat old mistakes. Client irritation rises for reasons nobody can quite pin down. Escalations become more frequent. The people who remain inherit more ambiguity and less context. Training gets thinner because the informal teachers are gone. Every new problem now takes longer to understand because the people who used to recognize it early are no longer in the room.

None of that shows up as neatly as salary savings.

AI Does Not Automatically Capture What Walks Out the Door

This is another place where the story gets too convenient.

Some leaders implicitly believe that if work has passed through systems long enough, and if AI can search, summarize, and generate effectively enough,

then
the organization has already captured most of what matters.

It has not.

AI can retrieve documented knowledge. It can synthesize artifacts. It can help
teams externalize more than they used to. All of that is useful. But AI does not automatically preserve the human context that was never fully encoded in
the first place.

It does not know why a client stopped trusting a certain reporting flow after a
bad quarter three years ago. It does not know which internal team always says
yes in a steering committee and then quietly blocks implementation later. It does not know why a technically elegant option will fail politically inside an account unless somebody in the room has lived that account long enough to say
so.

That is why reduction-first firms can become strangely confident at the exact
moment they are becoming more fragile. The dashboards improve. The visible work
arrives faster. The documentation may even get better because AI makes it cheaper to produce. Retrieval works nicely on the artifacts that were always written down.

What stays invisible is the silence of what was never written down in the first
place. Leaders see the system answering more quickly and conclude the system
knows more completely. Those are not the same thing.

That is not augmentation. It is overestimation.

Continuity Is a Growth Asset

One reason this chapter matters is that continuity is usually framed as a defensive concern.

Protect continuity.
Avoid disruption.
Retain institutional memory.

All of that is true. It is still only part of the picture.

Continuity is not only about preventing failure. It is also about enabling growth. You grow faster when the organization does not have to rediscover itself every quarter. You can expand offers faster when the people designing the next move still understand the last three moves. You can trust new capability more when it is being built on top of remembered judgment rather than rebuilt from fragments.

This matters especially in consulting, where trust compounds. A client does not
just buy output. A client buys confidence that your firm understands the terrain, remembers what matters, and will not make them pay twice for the same
lesson.

When firms cut too deeply, they do not just reduce cost. They often reduce the
quality of future growth.

The Beginning of the Do Not Cut List

This is where a practical rule starts to emerge.

Before leaders ask how much labor can be removed, they should ask which people
the firm cannot afford to lose even if AI has made part of their visible work faster.

That is the beginning of the Do Not Cut List.

Not a sentimental list. A strategic one.

Protect the people who carry client trust. Protect the people who translate between functions. Protect the people who teach others how the work really gets
done. Protect the people whose visible output understates their stabilizing effect on the system. Protect the apprenticeship nodes that create the next generation of judgment rather than only the current generation of throughput.

The list will vary by firm. The principle will not.

You do not cut the people who are helping the organization remember, adapt, and
cohere unless you are very sure you understand what will replace them.

Most firms are not nearly as sure as they think they are.

The next chapter is about the choice that follows from that realization. Once
you see what is at risk, cost management stops being a sufficient frame.

Do you use AI to substitute for people, or to multiply what the right people can do?

Chapter 4 — The Multiplier Model

By the end of the last chapter, the problem should feel clear.

AI changes delivery economics.
The easy win is reduction.
Reduction can quietly destroy continuity, trust, apprenticeship, and growth capacity.

That still leaves the real strategic question unresolved.

If cutting too deeply is the wrong answer, what is the right one?

This chapter offers the book's central fork.

I call it the Multiplier Model because the real decision is not whether AI gets
used. That part is already happening. The real decision is what kind of force a
firm wants AI to become inside the business.

Does it become a substitution engine?

Or does it become a multiplier?

One Tool, Two Roads

The easiest mistake leaders make at this stage is to talk about AI as if the tool itself determines the outcome.

It does not.

The same underlying capability can drive two very different strategies.

One firm uses AI to reduce the number of people required for the same basic work. It compresses labor, narrows teams, captures savings, and treats the visible efficiency as the primary value.

Another firm uses AI to increase what its people can handle, improve how work
flows, preserve the knowledge and judgment that matter most, and redirect the
freed capacity into deeper client work, new offers, faster learning, and new revenue.

Same pressure. Same class of tools. Different governing idea.

That is the fork.

The Substitution Road

The substitution road is the more intuitive one, especially in the middle of the dip.

The logic sounds disciplined. Use AI to reduce labor. Remove cost faster than
the market reprices you. Simplify the organization. Protect near-term margin.
Prove to clients and stakeholders that you are taking AI seriously.

There is a reason this road attracts smart leaders. It promises proof quickly. It creates visible movement. It looks decisive.

At first, it can even look superior.

The firm gets leaner. Some work gets faster. Utilization pressure eases. Cost
comes out. Executive narratives become easier to tell. The market may reward
the appearance of discipline long before it tests the deeper consequences.

That is why the substitution road can feel like the grown-up option.

It also creates a predictable set of downstream problems.

As labor is removed, the firm often weakens its developmental pipeline, thins the layer of people who carry cross-functional understanding, and reduces
the surface area from which higher-value capability could have grown. The remaining team may become more efficient and less adaptable at the same time.

The result is not always collapse. More often it is narrowing.

The firm gets better at defending a smaller future.

The Empowerment Road

The empowerment road starts from a different premise.

It assumes the productivity unlocked by AI is too valuable to spend only once.

Instead of asking only, "how much labor can we remove," this road asks which
people become more valuable when amplified, which parts of the workflow should
be redesigned rather than merely sped up, what knowledge and continuity

must be
preserved, where the newly freed capacity should go, and how the firm can convert productivity into stronger offers and stronger growth.

This road still cares about efficiency. It is not anti-cost and it is not anti-discipline. It treats efficiency as fuel rather than destination.

On the empowerment road, AI is used to multiply the reach of people who carry
judgment, trust, context, and learning capacity. It is used to make relationship owners more responsive, delivery leads more scalable, technical experts more leveraged, internal teachers more effective, and small firms more
capable than their headcount would once have allowed.

This is the road that turns productivity into capacity rather than only into savings.

Twelve Months Later

The fork becomes easier to understand when you stop looking at it as a belief
system and start looking at it over time.

Twelve months into the substitution road, a firm may look cleaner. Headcount is
lower. Some margins may be steadier. Some work is faster. But the internal conversation is often still dominated by cost, pressure, and replacement logic.
The firm has moved, but mostly by compression.

Twelve months into the empowerment road, a firm may look messier from the
outside because reinvestment is harder to narrate than cuts. But internally, something more interesting is starting to happen. The firm has new capability
nodes. More work can be handled by the same people without reducing them to
exhaustion. Patterns are being reused. New offers are emerging. Certain people
have become visibly more powerful because AI has increased their range.

The first road produces faster proof.

The second road produces stronger options.

Twenty-Four Months Later

Two years in, the difference widens.

On the substitution road, the firm is often fighting a more defensive battle than it expected. It still has efficiency stories, but those stories have become easier for competitors to imitate and easier for clients to demand as table stakes. The original savings are real, but less differentiating. If the firm has not built something above them, the business model keeps shrinking toward the more commoditized part of the market.

On the empowerment road, the firm has had time to turn reinvestment into shape.
This is where capability building, workflow redesign, better client conversations, and selective new offers start to compound. The gains are no longer only internal. They begin to show up in what the market can actually buy.

That is the difference between using AI to defend margin and using AI to expand
value.

Thirty-Six Months Later

Three years in, the firms do not merely look different. They are different.

One has become leaner, tighter, and more exposed to further repricing because
its main story is still that it can do familiar work with less labor.

The other has changed the kind of firm it is becoming. It may still deliver, but delivery is no longer the whole value story. It now sells more capability, more judgment, better leverage, better system design, better transfer, and better growth logic than it did before.

The first firm often says, "we adopted AI."

The second can say, "we became more capable because we reorganized the business
around what AI made possible."

That is a very different sentence.

Why Firms Default Left

Firms default left because substitution is easier to authorize and easier to explain. It has cleaner math, a shorter narrative, and a result that finance, the board, procurement, and an anxious market can all recognize quickly.

The multiplier road asks for something harder: a growth thesis. Leaders have to
protect certain people when the visible spreadsheet argues for savings, redesign

work instead of merely compressing it, and decide where the productivity should
go before the organization spends it accidentally.

A multiplier is not a pep talk or an abstract belief in people. It is a practical operating choice: keep the right people, redesign the workflow around
leverage, use AI to raise the range of trusted operators, reinvest capacity deliberately, and convert the gains into stronger offers, deeper client value, or faster growth.

That road is not softer than substitution. It is harder. It demands more discipline, more intentionality, and more managerial honesty. It just uses that discipline to build rather than merely to cut.

How This Applies to a Firm of One

The model works at smaller scale too.

If you are an independent consultant, the substitution road means using AI to
sell the same basic labor more cheaply, more quickly, or in more volume. You
may win some business that way. You may also train the market to experience you
as a lower-cost delivery engine with better tooling.

The empowerment road for a solo operator looks different, but the logic is the
same. Use AI to raise the level of judgment, synthesis, responsiveness, tailoring, and leverage you can bring to a client. Use it to expand the scope of what you can credibly handle, not just to make the old unit of work cheaper.

The question is still the same:

Are you using AI to substitute for labor, or to multiply value?

The Choice That Organizes the Rest of the Book

This chapter is not the full playbook. It is the decision that makes the playbook necessary.

Once you see the Multiplier Model clearly, the remaining questions start to organize themselves.

Who gets multiplied and who gets compressed?
Where does the productivity go?
What should be protected?

What should be redesigned?
What should the firm stop selling?
What should it start selling?

Those are the questions the rest of this book is trying to answer.

But the first answer has to be this one:

AI does not force one strategy on a consulting firm.

It forces a choice.

Chapter 5 — Who AI Replaces, Who It Multiplies

Once leaders see the Multiplier Model clearly, the next question becomes practical very quickly.

Who, exactly, belongs on which side of the fork?

That question is where many firms start sliding back into old habits. They agree in principle that the wrong cuts are dangerous. They agree that some people become more valuable when amplified. Then they walk into a talent review
and discover they still do not have a usable way to sort roles and people under
AI pressure.

So this chapter has one job:

turn the strategy into a talent decision tool.

The First Mistake: Treating All Roles as if AI Hits Them the Same Way

It does not.

Some roles are under direct compression. Some are changing shape but remain
essential. Some become more valuable because AI expands their range. Some are
quietly so important to continuity and judgment that cutting them on visible productivity alone is reckless.

If you do not distinguish those categories, AI decision-making gets sloppy very

fast. Everything starts to look like a cost problem, because cost is the one thing the spreadsheet can see clearly.

That is how firms end up cutting the wrong people and preserving the wrong structures.

Role Triage

The practical answer is a four-part role triage:

- Compress
- Augment
- Multiply
- Protect

The goal is not to make talent decisions automatic. The goal is to stop making
them blind.

Compress

Use Compress for work that is routine, bounded, repeatable, legible, and only
weakly tied to trust, continuity, or judgment.

This is where AI is most likely to reduce the human labor required substantially. Standard first-draft work belongs here. Some predictable analysis. Boilerplate-heavy delivery. Certain forms of documentation. Certain
forms of implementation support. Work whose value is mostly throughput.

That does not mean every person doing that work should be discarded. It means
the work itself is under compression, and the old staffing and pricing logic around it cannot be taken for granted.

This is where many firms make their second mistake. They correctly identify compressible work and then wrongly assume that everything and everyone attached
to that work is equally compressible too.

That is not triage. That is panic with a dashboard.

Augment

Use Augment for roles that still matter but can now operate with more range,
speed, or coverage because AI changes the workflow.

These are often people whose jobs still require judgment, but whose first pass,
preparation burden, or information synthesis can now be accelerated. Delivery
leads often belong here. Analysts. Solution architects. Account managers. People whose value does not disappear when AI enters the system, but whose
workflow absolutely should.

Take a delivery lead as an example. Before AI, that person may have spent large
amounts of time assembling status, chasing inputs, cleaning draft language, and
preparing for stakeholder conversations. After AI, the value is not just that the same person can do those tasks faster. The value is that the workflow can
be redesigned so more of that person's time goes into judgment, escalation, coaching, and keeping the account out of trouble.

Augmentation is where many leaders stop too early. They say, "good, this role
is safer than we feared," and move on.

The better question is: if this role can now operate with more leverage, what
should we redesign around that fact?

If you do not change scope, expectations, measures of value, and workflow, you
are using AI to create local efficiency without changing the system around it.

Multiply

Use Multiply for people whose value expands materially when AI increases their
reach.

These are not merely people who get faster. They are people who become more
powerful.

A high-context operator who can now synthesize more options before a client
meeting. A cross-functional translator who can turn AI into better alignment instead of just faster output. A trusted expert who can scale judgment across
more teams. A capability builder who can codify patterns, teach them, and

reuse
them faster than before.

Multipliers are where the book's title starts to matter most.

If your best people become more valuable with leverage, then productivity should not be treated only as a labor-reduction event. It should also be treated as an opportunity to expand scope, deepen client value, and increase the rate at which the firm learns and compounds.

The multiplier question is not:

Can this person do the same work faster?

It is:

What new value becomes available because this person now has leverage?

Protect

Use Protect for people whose visible output understates their strategic value.

These are often Knowledge Carriers.

They hold continuity. They stabilize relationships. They connect functions that
otherwise misread one another. They prevent expensive mistakes before those
mistakes become visible. They train others informally. They know why the system
looks the way it does, not just what the documentation says.

Some firms will resist this category because it sounds too cautious.

It is not caution. It is risk management.

The Protect category exists because some people are easy to underestimate right
before the organization needs exactly what they carry.

That is why this chapter's tool is not just role triage. It is also the beginning of the Do Not Cut List.

The Role Triage Framework

The simplest way to use the model is to ask seven questions:

1. Is the value mainly in throughput, or in judgment?
2. How much of the role is routine and legible?
3. How much of the role depends on trust, continuity, or translation?
4. If this person left tomorrow, what would break first?
5. Would AI make this person cheaper, or more powerful?

6. Is the work compressing, or is the person becoming more leverageable?
7. Are we measuring visible output while missing carried value?

Those questions are now captured in the Role Triage Framework in the appendix.

For the book itself, what matters is simpler: this chapter should be usable in
a talent review next week, not just memorable in a notebook.

Warning Signs You Are About to Misclassify Someone

You are likely about to get a role wrong if it looks redundant only because its
best work appears as fewer problems.

You are also likely to misclassify a role if it sits in the middle. Middle layers are easy to disrespect in periods of efficiency pressure because their value often looks like translation, clarification, exception handling, or keeping the machine from rattling itself apart.

That work does not always look heroic. It looks ordinary right up until it is gone.

Another warning sign is interruption visibility. If a role has weak dashboard visibility but everyone gets nervous when that person is unavailable, the firm
should slow down before treating it as redundant.

That is often a Protect or Multiply signal masquerading as an efficiency candidate.

Why Seniority Is a Bad Shortcut

One of the laziest ways to sort talent after AI is by seniority.

Keep the senior people.
Automate the junior work.
Shrink the middle.

Sometimes that logic will partly map to reality. Often it will miss the actual distribution of value.

Some junior work is clearly compressible. Some senior work is expensive theater
wrapped around familiar labor. Some middle-layer people are carrying more strategic continuity than anyone notices because their job title sounds ordinary.

If you use seniority as a proxy for irreplaceability, you will protect the wrong things.

The right shortcut is not hierarchy.

It is value shape.

The First Pass on the Do Not Cut List

Before a firm reduces any role under AI pressure, it should pause and ask whether that role or person does any of the following:

- owns key client trust
- translates across business and technical teams
- carries undocumented context
- acts as an apprenticeship node
- catches expensive mistakes early
- holds continuity across multiple moving parts

If two or more are true, the default should not be Compress.

The default should be:

stop and look harder.

That does not mean "never change the role." It means do not let visible productivity alone drive a decision that may be cutting hidden value.

By this point, a consulting leader should be able to do three things immediately: sort roles and people into provisional buckets, run a first-pass Do Not Cut screen before reduction decisions are made, and spot where visible
productivity is disguising hidden continuity value.

That is the first true playbook move in the manuscript, but it still leaves a deeper question unresolved. Some people are clearly becoming more valuable now,
but why are they becoming more valuable? What capabilities are actually rising
in value?

That is the next chapter.

Chapter 6 — The People You Didn't Know You Had

Some firms are going to misplay AI by cutting too deeply. Other firms are going

to misplay it in a quieter way: they are going to keep the right people and still underuse them. That may sound like a smaller mistake. It can become just
as expensive.

The question at this point is no longer only who should we not cut. It is also
who have we failed to notice.

Most Organizations Know Less About Their People Than They Think

Organizations are usually better at measuring current output than future range.

They know who closes the ticket, ships the code, runs the meeting, fills the deck, and manages the account. They know who is visibly productive in the role
they already have.

What they often do not know is which people would look different if the role changed, the friction dropped, or the person suddenly had more leverage than
the organization has ever given them before.

That is one reason AI creates such a strange management moment.

It does not only compress obvious work. It also reveals people who were more
capable than the system had required them to be.

Some of the biggest AI winners inside an organization will not be the people everyone already assumes are stars. They will be the people whose judgment, curiosity, synthesis, taste, or translation ability has been trapped inside narrow workflows and repetitive labor.

Those are the people you did not know you had.

Underperformance and Underexpression Are Not the Same Thing

This is the distinction most firms still miss.

A person can underperform because they lack the judgment, discipline, or skill
for the role.

A person can also look ordinary because the role is asking too little of what they are actually capable of.

Those are not the same problem, and AI makes the difference more important.

When repetitive burden drops, when first drafts are easier, when synthesis is
faster, when research becomes cheaper, and when a person can externalize more
of their thinking with better support, some people do not merely speed up. They become legible in a new way.

That is not promotion theater. It is latent capability becoming visible.

What Latent Capability Looks Like

It rarely announces itself with a job title.

More often, it shows up in signals:

- a person solves better problems than their role requires
- peers trust their judgment even when the hierarchy does not highlight them
- they ask unusually good questions across domains
- they improve fast once given more context
- they can explain complexity clearly
- they seem more constrained by workflow than by ability

Those signals matter because many firms still evaluate people through the narrow aperture of the work they are currently assigned.

That was always limiting. AI makes it more expensive.

If leverage expands what a person can do, then the organization has to get better at spotting who might become more valuable once friction, repetition, and bottlenecks are reduced.

A Different Kind of Talent Review

The usual talent review asks:

- who is performing well now?
- who is ready for more responsibility?
- who is at risk?

Those are still useful questions.

They are no longer enough.

An AI-era talent review also has to ask:

- who is underused?
- who gets stronger when the problem gets messier?

- who is judged mostly on throughput but might create more value in synthesis,
  translation, redesign, or judgment?
- who becomes more interesting when repetitive work is removed?

That is the basis of a latent-capability review.

And if it is not built into the operating rhythm of the firm, the organization will keep making one of the most expensive mistakes in the AI era:

seeing only the capability it already knows how to buy from its own people.

The Hidden Upside of AI

This is where the conversation usually gets too narrow.

Most organizations are still talking about AI as if its primary job is to reduce effort on known work.

That is part of the story.

The more interesting part is that AI can reduce the distance between what a
person can see and what they can produce.

That matters in consulting because so much value comes from people who can
frame problems, connect domains, synthesize messy inputs, and create useful artifacts quickly enough for the business to act on them.

A person who once looked like a solid analyst may become a much stronger designer of client-facing thinking when AI removes the drag around drafting and
research. A delivery manager may become much more valuable as a risk spotter and
workflow improver once status assembly stops consuming so much attention. A
quiet operator may emerge as a force multiplier because AI finally lets them externalize what they have been carrying internally for years.

This is already happening in individual practice all over the place. The organizational version is simply slower to recognize itself.

From Recognition to Activation

This is where many firms stall. They notice someone is stronger than the role
has been showing, and then they do nothing with that realization. The daily
work keeps moving. The org chart stays the same. Managers tell themselves

they
will create more room later.

Later rarely comes on its own.

Seeing latent capability is not enough. The firm has to activate it. That means
leaders need a repeatable way to surface underused people, test for expanded range, and create safer first opportunities for them to operate with more leverage. The goal is not to throw people into bigger roles and hope. It is to run better experiments.

A simple first pass looks like this:

1. List 10 to 15 people below the obvious senior stars.
2. Mark who is currently judged mostly on throughput.
3. Mark who shows trust, synthesis, translation, or unusual judgment.
4. Ask where AI could remove repetitive burden from each person.
5. Choose 2 to 3 people to test in broader, AI-amplified work.

The point is not to identify hidden geniuses. It is to stop assuming the organization already knows the upper range of the people it has. What matters
is that a leader can run this in a real talent conversation without turning the
meeting into theater.

Safe Activation Tests

The best activation tests are small, real, and slightly above the current role.

Ask someone to solve a problem one layer above their normal scope. Give them AI
support and better context. Let them synthesize across functions. Let them prototype a workflow improvement. Let them build a stronger first pass on a
client-facing artifact than their title would normally authorize.

Then watch what happens.

Do they get clearer as the problem gets bigger?
Do they create value faster once friction is reduced?
Do they show more judgment than the old role ever required?

Those are multiplier signals.

The wrong activation test is simply to give someone more of the same work and
call that growth. That is not activation. That is load.

What This Means for a Firm of One

This chapter also matters for the independent consultant.

A solo practice can underuse itself just as easily as a larger firm can underuse its people. If you only use AI to do the work you already sell a little faster, you may never discover the next layer of value you could bring.

The latent-capability question for an independent is different in wording but not in spirit:

What part of my own range has been trapped inside the way I currently make
money?

That is how a delivery specialist becomes a strategist, a builder becomes a designer of systems, or a subject-matter expert becomes a more valuable guide
because AI reduces the friction around expression, synthesis, and iteration.

The same principle holds:

You may be underestimating your own leverage because the current role is too
small.

The old management habit is to ask whether a person is succeeding in the role
they already have.

The better question in the AI era is whether the role is still large enough for what the person could become with leverage.

That is not a sentimental question. It is a strategic one.

Firms that learn to spot underexpression will find capability their competitors never realize they are leaving unused. Firms that only optimize what is already visible will keep shrinking talent to fit the old boxes and then wonder why growth feels so hard to generate from inside.

By this point, a consulting leader should be able to identify people who may be
underused rather than underpowered, run a first-pass latent-capability discovery exercise, and choose a few safe activation tests instead of making assumptions from job titles alone.

That is the second playbook move in the manuscript. It also sets up the next
question.

Once you know who may become more valuable with leverage, you still have to

decide where that leverage should go.

That is the chapter after this one.

Chapter 7 — Every Engagement Needs a Growth Thesis

AI creates one of the most dangerous illusions in consulting. It can make a firm feel smarter the moment it starts moving faster, even if nobody has decided where the new capacity is supposed to go.

That is not strategy. It is acceleration without destination.

This is where many AI engagements quietly fail, even when the work itself is
good. The workflow improves. The drafts come faster. The turnaround time drops.
Some labor is saved. Everyone involved can point to visible gains. But when the
engagement ends, the firm is still left with the most important unanswered question:

Where does the productivity go?

If the answer is vague, the gain usually disappears. It gets absorbed into busyness, local relief, or margin smoothing. People feel pressure ease for a while. The organization congratulates itself. Then the old growth constraints return, because the productivity was never given a strategic destination.

That is why every meaningful AI engagement needs a growth thesis.

A Growth Thesis Is More Than a Positive Attitude

This phrase can sound softer than it is, so it helps to make it plain.

A growth thesis is a specific answer to the question of what new value the organization intends to create with the capacity AI frees up.

Not what it hopes might happen.

What it is choosing to pursue.

That choice can take several forms. The firm may decide to produce more output
against existing demand. It may use the capacity to go deeper with clients and
move upstream into higher-value work. It may convert repeated client need

into
new services. It may use the gain to enter adjacent markets. Or it may invest
the capacity internally so the firm can actually sustain the shift instead of borrowing against its own future again.

Those are different choices. They require different owners, different measures, and different sacrifices. That is exactly why they have to be chosen on purpose.

The Growth Allocation Decision

The practical move in this chapter is simple: when an AI initiative creates capacity, decide explicitly where that capacity goes.

Five allocation paths matter most:

1. more output
2. deeper client work
3. new services
4. new markets
5. internal capability

The categories are straightforward. The discipline is not.

Most firms never really make this decision. They talk in broad terms about growth, then let the capacity get reclaimed by the nearest demand, the loudest sponsor, or the oldest work pattern still sitting in the system. The result is predictable. AI makes the firm busier, but not meaningfully stronger.

The better move is to force a short sequence of questions:

1. What capacity has actually been created?
2. Is that gain repeatable or temporary?
3. Which growth constraint matters most right now?
4. Which allocation path best addresses that constraint?
5. What work will stop so this choice becomes real?
6. Who owns the outcome?
7. What will we measure ninety days from now?

That is the Growth Allocation Decision.

The point is not to turn judgment into a formula. It is to stop letting productivity gains vanish into the background noise of the firm.

I have seen how badly this goes when the growth question never gets asked. I
walked into a room once where fifteen people had built the wrong thing for understandable reasons. They had taken the data they had and built what

amounted to prettier charts. The real business question had never been asked.
So I told them, as directly as I could, that I had held their jobs before and made the same mistake myself. They had started with the available data instead
of with the people who actually made the money.

The room got very quiet. Then the leader at the table smacked it hard and said,
"That's why we brought you here. That's why we are paying your rate. We needed
to hear that."

That is what a growth thesis often sounds like when it first enters the room. Not "how can we make prettier output from the data we already have?" but "how
do the people doing the work actually create value, and what gets in the way?"
Before a firm scopes an AI engagement, someone should spend real time with the
end users who create the business result. The growth thesis lives there, not in
the dashboard.

Why Productivity So Often Disappears

The failure mode here is familiar. A team saves time, nobody removes anything
from the workload, and the new capacity gets reclaimed by the old system.

Cuts can be announced in a sentence. Growth has to be allocated. It needs an
owner, tradeoffs, and room to become real. Without that discipline, AI simply
raises expectations around existing work and calls the result progress.

How to Choose the Right Allocation Path

The best allocation path depends on the constraint the firm is actually facing.

If demand already exists and delivery is the bottleneck, more output may be the
right answer. If the firm already has trust with clients but not enough share of the problem space, deeper client work may matter more. If repeated needs are
showing up across accounts, the better move may be to turn that pattern

into a
new service. If the firm has built portable capability, it may be time to test a new market. And if the firm keeps talking about growth while running on fragile workflows, unclear methods, and uneven capability, internal investment
may be the real priority.

What matters is that the choice match the constraint.

This is where many firms drift into self-deception. They choose the allocation category that sounds most ambitious rather than the one their system
can actually support. A new market sounds exciting. A new service sounds innovative. But if the real problem is that the current firm cannot reliably scale judgment, preserve knowledge, or train people into the new way of working, the most strategic use of the capacity may be internal.

That is not retreat. It is preparation that pays.

Signals That the Allocation Is Working

The allocation is probably working if the new capacity becomes visible in one
of three ways fairly quickly.

First, the chosen path starts changing what the firm can now do, not just how
fast it does familiar work. Second, someone can point to clear ownership rather
than saying the gains are being "used across the team." Third, the new allocation begins producing evidence that the firm is moving toward a stronger
future: deeper client dependence, a sharper offer, more resilient delivery, or better ability to grow without hollowing itself out.

The failure signals are just as important.

If no one can say where the productivity went, it is gone. If utilization went up but strategic position did not, the capacity was consumed rather than invested. If managers quietly refill the time with old work, the system has chosen maintenance over growth. If the gain is celebrated once and then disappears into the baseline, the firm never made an allocation decision at all.

A Better Standard for AI Engagements

This changes how a consulting firm should scope its own work.

An AI engagement should not be considered fully designed until someone can
answer three questions clearly:

- What capacity will this create?
- Where will that capacity go?
- How will we know it actually went there?

If the sponsor cannot answer those questions, the engagement may still produce
efficiency. It does not yet have a growth thesis.

That distinction matters because a firm that delivers efficiency without a destination is helping the client optimize the present. A firm that helps the client allocate new capacity is helping design the future.

That is a different level of consulting.

What This Means for a Firm of One

The same choice exists for an independent consultant.

When AI makes your current work faster, the extra capacity can disappear just as
easily as it does in a larger firm. You can fill it with more of the same delivery. You can let clients absorb it through lower prices and tighter turn times. Or you can decide, deliberately, what that freed range is for.

For a solo operator, the allocation paths still hold. More output may mean taking on more of the right work. Deeper client work may mean moving from
execution into diagnosis or design. New services may mean packaging a pattern
you were previously just performing ad hoc. New markets may mean finally taking
an offer into a different segment. Internal capability may mean building the method, artifacts, and workflow discipline that let you stop selling yourself only as labor.

The core question is the same:

What is this new capacity for?

If you do not answer it, the market will answer it for you, and it will usually answer in favor of cheaper delivery.

The old habit is to celebrate the gain and move on.

The better habit is to treat every meaningful productivity gain as an allocation decision.

That forces a firmer kind of leadership. It requires saying no to some work so another kind of work can actually grow. It requires naming the constraint that matters most. It requires accepting that more speed is not yet
the same thing as more strategy.

By this point, a consulting leader should be able to reject AI initiatives with no strategic destination, choose a primary path for freed capacity, and assign both ownership and a near-term success signal to that choice.

That is the third playbook move in the manuscript.

The next question is not where the capacity goes. It is what kind of work the
firm should stop relying on in the first place.

Chapter 8 — Stop Selling the Wrong Work

By this point in the book, the question is no longer whether AI changes consulting economics. It does. The question is whether the firm is still trying
to build its future on work that is getting easier to compare, easier to speed up, and harder to defend as premium.

That is where many firms now are.

They have better tools. They may even have sharper people. But they are still
trying to protect revenue streams whose main story is visible effort. When that
is the story, AI does not just improve the work. It weakens the pricing logic around the work.

This chapter is about how to stop relying on the wrong work before the market
forces the decision for you.

Busy Work Can Still Be the Wrong Work

One of the hardest things for a consulting firm to admit is that a service line
can still be active, profitable, and important to current revenue while also becoming a weaker foundation for the future.

That is the trap.

The wrong work is not always bad work. It is often familiar work. It may still
help clients. It may still keep teams busy. It may still throw off margin. But
if the offer is becoming easier to compare on price, easier to accelerate with AI, and harder to explain as premium judgment, it is no longer the kind of work a firm should quietly bet its identity on.

This is where consulting leaders have to separate current revenue from future reliance.

Those are not the same thing.

What Makes Work Harder to Defend

The pressure usually shows up before firms admit what it means.

Clients start asking why a piece of work still costs what it used to cost now that drafting, synthesis, coding, documentation, research, or workflow setup can be done faster. Sponsors become less patient with teams whose main value is
that they can put more people on the problem. Procurement gets more confident.
Internal buyers start wondering whether some of the work can be done by a smaller team, a cheaper provider, or their own people with better tools.

None of that means the work has no value.

It means the old pricing story is weakening.

That is the difference firms have to see clearly. The market does not need to believe the work is worthless. It only needs to become less convinced that the
old premium still makes sense.

The Service Portfolio Triage

The practical move in this chapter is to sort the current portfolio into four buckets:

1. Defend
2. Redesign
3. Sunset
4. Climb Above

The names are straightforward. The discipline is not.

Defend means the offer is still genuinely differentiated. The client is buying more than labor. The work still connects to trust, judgment, outcome ownership, or access to problems the client cannot easily solve alone.

Redesign means the need is still real, but the offer shape is wrong. The firm should keep the problem space while changing scope, deliverables, pricing, or
the way value is explained.

Sunset means the offer is increasingly commodity, increasingly price-exposed, and no longer strategic enough to anchor the future. It may continue for a while. It should not keep owning the firm's attention.

Climb Above means the current work still matters, but the more defensible position is upstream. The firm should move from delivery into diagnosis, redesign, enablement, governance, capability building, or some other form of
higher-order problem ownership.

One practical distinction helps here. Redesign keeps you in essentially the same problem space but changes how the value is packaged and delivered. Climb
Above uses the current work as proof and entry, then moves the client into a different level of problem ownership. If redesign is about reshaping the offer, climb above is about changing the altitude of the conversation.

The key is to use these buckets on current revenue, not on imaginary offerings that only exist in the partner retreat deck.

The Questions That Sort the Work

For each major offer or service line, ask:

1. Is the client primarily buying judgment or labor?
2. Has AI reduced the visible effort in a way the client can now notice?
3. Is the offer easy to compare against cheaper alternatives?
4. Does the work create trust that leads to higher-value follow-on work?
5. Are we being paid for outcomes, direction, or mostly for staffed effort?
6. If this offer shrank by thirty percent, would we still want to build the future of the firm around it?
7. What higher-value work could this offer lead into if repositioned well?

Those questions do two things at once. They expose which work is becoming hard
to defend, and they keep the firm from confusing current busyness with strategic
value.

That matters because many firms are still treating a full pipeline as proof that the portfolio is healthy. It is not. A full pipeline can simply mean the market has not repriced the work all the way yet.

The Wrong Work Usually Has Visible Symptoms

Leaders rarely need a spreadsheet to know where the pressure is gathering. The signals are usually already in the room.

Pricing is still tied mostly to hours or headcount. Clients start asking why the work should still cost this much. Teams spend more time defending effort than describing outcomes. AI mainly makes the work faster rather than more valuable. The offer stays busy but produces little reusable method, little relationship depth, and little follow-on leverage for the rest of the firm.

That is the wrong work warning you before the income statement does.

The mistake is to answer those signals with better rhetoric around the same offer. The stronger move is to decide whether the work should be defended, redesigned, sunsetted, or climbed above.

Do Not Confuse Delivery Importance with Business Priority

Some work still matters deeply to the client and still should not define the future of the firm.

That sentence is hard for consulting leaders because many firms were built by doing necessary delivery work well for a long time. There is pride in that. There should be.

But necessity is not the same thing as defensibility.

A firm may still need to do implementation work, migration work, or execution support. The question is whether that work is the business or whether it is now the doorway into a more valuable business.

That is where the distinction between Redesign and Climb Above becomes so important. Some offers should be reshaped. Others should become the entry point to a more strategic conversation. If the client still needs the same work but the current packaging is getting weak, redesign it. If the client is really buying access to a bigger problem that the current work only reveals, climb above it. Very few offers should remain untouched just because they are familiar.

What a Smaller Firm Should Do First

Mid-market and boutique firms do not have the luxury of redesigning everything
at once.

That is why portfolio triage matters.

The first move is not to launch five new offers. It is to identify one revenue line the firm is increasingly uneasy defending and decide whether it should be
redesigned, sunsetted over time, or climbed above.

Then pick one adjacent, more defensible version of the value already visible in
that work.

If a firm is known for implementation, the next offer may be workflow redesign.
If it is known for technical delivery, the next offer may be AI-enabled operating design, capability transfer, or governance. If it is known for staffed execution, the next offer may be decision acceleration, enablement, or
problem diagnosis that reduces the amount of brute-force labor the client needs
to buy.

This is not reinvention by slogan. It is portfolio movement by sequence.

What This Means for a Firm of One

The same trap exists for an independent consultant.

A solo operator can stay fully booked selling work the market is steadily teaching them to price lower. That is one of the easiest ways to get trapped by
your own competence.

The same four buckets apply.

Which parts of your work are still genuinely defendable? Which need redesign?
Which should be allowed to taper off instead of continuing to anchor your identity? Which pieces of current client work are really a doorway into more valuable advisory, design, or capability-building work if you are willing to name them that way?

That is how a one-person practice stops being defined only by what it can deliver directly and starts being valued for how it helps clients think, decide, and build differently.

The old habit is to ask whether a service line is still producing revenue.

The better habit is to ask whether that revenue is training the firm toward the
future or keeping it dependent on a past the market is already repricing.

That is not a theoretical question. It is a portfolio decision.

By this point, a consulting leader should be able to sort the current service portfolio into Defend, Redesign, Sunset, and Climb Above, identify which revenue lines are most exposed to delivery repricing, and choose where the firm should stop relying on visible effort as the business.

That is the fourth playbook move in the manuscript.

Necessity is not the same thing as defensibility.

The firms that learn that in time get to redesign themselves on purpose. The ones that do not usually get redesigned by the market.

Chapter 9 — Build the New Service Portfolio

Once a firm can see the wrong work clearly, the natural question is what replaces it.

That is where many firms become vague again.

They say they want to move upstream. They say they want to sell more strategy.
They say they want to become more advisory. All of that may be directionally true. None of it is a service portfolio.

The next step has to be more concrete than that.

The New Portfolio Cannot Be Built from Slogans

A firm does not escape delivery pressure by announcing a more elevated identity.
It escapes by packaging work clients will still pay premium prices for when the
old delivery story gets weaker.

That usually means moving toward offers built around capability, redesign, decision quality, knowledge transfer, and operating leverage rather than around
visible effort alone.

The shift matters because AI changes the shape of the output. It does not just

make old work faster. In the hands of someone who knows what they are doing, it
can make one person capable of producing work that used to require a small team
or much more elapsed time.

I have seen that up close. I was added to a GitHub Copilot pilot even though I
was already experienced, which meant the tool was not teaching me the work. It
was amplifying work I already knew how to do. On the first day I had meaningful
leverage from it inside a constrained enterprise environment, I was not
participating in an ordinary productivity pilot. I walked into the piano store
and played the Rach 2.

In roughly two hours, I produced what would normally have been a full day's
work and a persuasive case study for the pilot. Two hours later, I delivered
another day's work and a second, different case study. The client leaders, the
AI sponsors, and the executive backing the effort were blown away, not because
AI had replaced expertise, but because it had suddenly multiplied it in public.
That credibility became the foundation for a much bigger strategic proposal a
couple of weeks later.

That story matters here because it changes the portfolio question. If one
capable practitioner can now produce far more strategic output, then the firm
should not only ask how to price the old work. It should ask what new offer
shape becomes possible because the old effort bottleneck has changed.

Five Replacement Offer Types

The practical move in this chapter is to stop talking about "more strategic
services" in the abstract and start designing concrete replacement offers.

Five categories are especially useful:

1. capability building
2. workflow redesign
3. knowledge preservation
4. human-AI operating design
5. decision acceleration

These are not the only possible offers. They are the most natural starting points for firms trying to climb above labor-heavy delivery without pretending
they have become a strategy house overnight.

Capability building means helping the client's own people become more effective
with new tools, methods, and patterns. The buyer is usually trying to build internal strength, not just finish a project.

Workflow redesign means changing how work moves, where decisions happen, and
how handoffs are structured. The deliverable is not just a faster step. It is a better operating flow.

Knowledge preservation means protecting continuity, institutional memory, and
role-critical judgment before it walks out the door or gets fragmented by AI-enabled change. The client is buying resilience, not just documentation.

Human-AI operating design means helping the client decide how people, tools,
workflows, and governance fit together. The value is in making AI dependable inside a real operating system rather than impressive in a demo.

Decision acceleration means helping the client move faster on higher-quality choices, not just produce more artifacts. The client is buying clearer judgment
under pressure.

Those are offers clients can buy because they solve real problems the old model
is getting worse at naming.

Build from Existing Proof, Not Fantasy

The easiest way to design bad new offers is to invent them from branding language rather than from work the firm has already proven it can do.

The better move is to build from adjacent proof.

That means asking seven questions for each candidate offer:

1. What recurring client problem does this solve?
2. Why is the current delivery-heavy version getting weaker?
3. What outcome will the client actually buy?
4. What judgment, method, or capability makes this offer defensible?
5. What proof do we already have from current work?
6. What existing service line does this grow out of?

7. What is the lightest viable version we could sell in the next ninety days?

This is one reason the old work still matters. Delivery should not always stay
the business. It can still serve as proof, entry, and pattern source for the next business.

That is what Climb Above was trying to protect in the last chapter.

What a Better Offer Actually Looks Like

Most firms do not need a magical new category. They need a sharper offer shape.

Every replacement offer should be defined in five parts:

- the problem
- the outcome
- the method
- the deliverables
- the next logical follow-on work

That sounds simple because it is. But many consulting offers are still too weak
on the middle three.

They know the problem. They know the deliverables. They are much less clear on
the outcome the client is truly buying, the method that makes the offer defensible, and the follow-on work the offer should naturally create if it is well designed.

That weakness becomes much more expensive when AI compresses visible effort.

If the offer is basically "we do the work for you with skilled people," then the method is not very visible, the follow-on logic is weak, and the outcome is
too easy to compare with cheaper alternatives. If the offer is "we help your people do this differently, preserve what matters, and build a more capable operating system around the work," the conversation changes.

One Adjacent Move Beats Five Grand Ambitions

This is where firms often overcorrect.

They realize the current portfolio is under pressure and respond by trying to
launch an entirely new consulting identity all at once. That move is usually

too big, too vague, and too detached from what the market already trusts them
to do.

The better move is one adjacent offer.

Pick one client problem the current work keeps exposing. Package the higher-value response to that problem. Sell the lightest viable version first. Use the current delivery relationship as proof of relevance, not as the final form of the business.

That is the sequence smaller firms need.

If a team has repeatedly been helping clients clean up inefficient workflow through implementation work, the adjacent offer may be workflow redesign. If it
keeps rescuing fragile knowledge handoffs, the adjacent offer may be knowledge
preservation and transfer. If it is repeatedly being pulled into strategic judgment calls late in the engagement, the adjacent offer may be decision acceleration or operating design.

The point is not to invent something exotic. It is to notice what the current work is already proving the firm should be selling more explicitly.

What This Means for a Firm of One

This chapter matters just as much for an independent consultant.

A one-person practice may not have a portfolio in the formal sense, but it still has an offer mix, a market story, and a default way of making money. The
same replacement logic applies.

What recurring problem do clients keep revealing underneath the work they hire
you for? What method do you already have that is more defensible than the labor
you are currently billing? What light version of a stronger offer could you sell in the next ninety days without pretending to be a different company than
you are?

That is how a solo operator moves from “I can do this work for you” to “I can
help you become more capable at this class of problem.”

That is a better business.

The old habit is to ask what services the firm currently provides.

The better habit is to ask what capabilities, methods, and problem ownership the firm has already earned the right to package more explicitly.

That is a portfolio-building question, not a branding exercise.

By this point, a consulting leader should be able to sketch two or three replacement offers for the current firm, explain why those offers are more defensible than labor-heavy delivery, and choose one offer to prototype first.

That is the fifth playbook move in the manuscript.

The new offer is only half the move. If it is still sold, staffed, scoped, and measured through the old delivery model, the old logic will pull it back down.

Chapter 10 — Redesign the Engagement

A better offer delivered through the old engagement model will usually end up
becoming the old offer again.

That is the problem this chapter has to solve.

This is where many firms quietly lose the future they have just started describing.

The Old Engagement Model Pulls Everything Back Down

The old model is familiar because it worked for a long time.

Sell hours. Scope tasks. Staff the work. Protect utilization. Show progress through visible activity. Keep the method mostly inside the firm. Let the client depend on the team for repetition. Call the engagement successful when
the deliverables are complete and the invoice is clear.

That structure can still feel normal even when the offer sitting on top of it has changed.

A firm may say it is now selling workflow redesign, capability building, or decision acceleration. But if the engagement still behaves like labor-heavy execution with a nicer title, the old gravity wins. The client experiences staffed effort. The team defaults to volume. The knowledge transfer gets thin.
The new offer gets dragged back down into the older business model.

That is why engagement redesign is not polish. It is structural.

What Has to Change

The practical move in this chapter is to redesign the engagement around five
shifts:

1. outcomes instead of effort
2. transfer instead of dependency
3. leverage instead of labor
4. method instead of heroics
5. follow-on value instead of one-off completion

Those shifts do not eliminate direct delivery. They make it more deliberate.

An engagement still needs work done. The question is what kind of work the firm
should do directly, what should be transferred into the client organization, what should be amplified through AI, and what should be measured so the client
ends stronger than it started.

That is a very different design problem from simply staffing the project well.

Start with the Real Buying Decision

Before the engagement is scoped, a firm needs to answer a blunt question:

What is the client actually buying?

Effort is one answer. Output is a slightly better one. But the stronger answers
usually sound different: capability, a better operating result, faster and clearer decisions, less fragility, stronger workflows, preserved knowledge, or a more reliable internal system.

That distinction matters because the engagement will follow the thing the buyer
really believes they are paying for.

If the client thinks they are buying effort, they will count people and hours. If they think they are buying output, they will count deliverables. If they think they are buying stronger capability or a better operating result, the engagement opens room for transfer, redesign, and leverage.

This is why new offers often need new buying language before they need new
delivery mechanics.

Scope the Direct Work on Purpose

The old model quietly assumes the firm should do as much of the work as possible because that is what makes the economics work.

The new model has to be more disciplined.

Some work still belongs in the firm's hands because it requires concentrated judgment, difficult synthesis, or the ability to move quickly across a messy problem. Some work should be done jointly because the point is not just to solve the problem once but to make the client more capable of solving it again.
Some work should be transferred because the future value lies in the client's own operating capacity, not in preserving dependence on the consulting team.

That is why one of the most useful engagement questions is:

What part of the work should we still do directly, and what part should the client be stronger at by the end?

That question changes scope immediately.

It also changes who has to be in the room. My neighbor of more than three decades is a retail consultant. He can walk into a midsize retailer and diagnose where they are wasting money because he knows how that business actually works. I can walk into a Snowflake environment and optimize warehouse
spend because I know that system deeply. Both kinds of expertise matter. The
retailer who hires only my neighbor gets strong diagnosis and weak technical execution. The company that hires only me gets optimized spend on a warehouse
full of the wrong priorities. The engagement that creates value needs both.

That lesson got into me much earlier than most consulting frameworks did. At
AT&T, during a strike, management had to go do union work. I spent two weeks in
a call center and discovered a painful manual process. I built a simple spreadsheet that they loved. But the spreadsheet was not really the insight. The two weeks doing the work was the insight. If I had not been sent there, I
would never have known what was actually worth improving.

You cannot meaningfully improve what you have never done. That is why engagement redesign has to include proximity to the real work and the people
who understand its pain from the inside, not just the people who know how

to
instrument it from above.

Use AI for Leverage, Not Theater

AI belongs inside the engagement, but not as a stage prop.

The point is not to impress the client by showing that the team used AI. The
point is to use AI where it creates leverage: faster first passes, better synthesis, stronger pattern recognition, clearer options, more reusable methods, lighter administrative drag, and better support for the people whose
judgment matters most.

When that leverage is real, the firm should not hide it and then keep billing as if nothing changed. It should redesign the engagement around the value that
the leverage now makes possible.

That often means fewer people in some places, more explicit transfer in others,
and clearer explanation of the method so the client understands why the work is
still valuable even though parts of it now move faster.

The old model sold labor hidden inside execution. The redesigned model sells
leverage visible inside a better outcome.

Rewrite Success Before the Work Starts

One of the easiest ways to sabotage a redesigned engagement is to keep the old
success criteria.

If success is still mostly measured by hours burned, deliverables completed, or
the appearance of constant team activity, the engagement will naturally drift
back toward the old model.

The better move is to rewrite success before the work starts.

A redesigned engagement should ask:

- what will be true at the end that is not true now?
- what will the client be able to do better on its own?

- what knowledge, method, or judgment will remain after the team leaves?
- what follow-on value should this work create?

Those questions move the measurement system from completion to strengthening.

That is not softer. It is more demanding.

I have seen this logic matter in a real constrained setting. A couple of weeks after establishing credibility in an enterprise AI pilot, I put forward what I called the Lewis and Clark expedition proposal. The point was not just to ask
for permission to use AI more broadly. It was to redesign the engagement before
the work started. The proposal framed the effort as controlled exploration with
clear stages, defined outputs, explicit learning transfer, and a path from early proof to something the organization could actually scale.

That is the pattern this chapter is trying to protect. The work was not scoped
as "give me more time with better tools." It was scoped around what the client
would learn, what would be demonstrated, how progress would be judged, and what
larger follow-on value the pilot could create. In other words, the engagement logic changed with the capability. That is what leaders need to do more often:
rewrite the shape of the work before the old model quietly rewrites it for them.

This is not abstract. It changes how proposals are written, how statements of
work are scoped, how teams are staffed, how status is discussed, and how final
success gets judged.

What This Means for a Firm of One

The same issue shows up in independent consulting.

A solo operator can build a stronger offer and still deliver it through a model that quietly says, "I am still mostly selling my time." That is how many good
independent offers stay trapped below their real value.

The redesign questions still apply.

What are you actually being bought for? What should the client be stronger at
after working with you? What part of your method can become visible and transferable? What follow-on value should the engagement create? Where does AI
let you create leverage that should change the shape of the work, not just your
speed inside it?

Those are not big-firm questions. They are consulting questions.

Deliver the New Offer Through a New Logic

The old habit is to improve the offer and leave the engagement alone.

The better habit is to redesign both at the same time.

That is what keeps the business model from snapping back to its default shape.

By this point, a consulting leader should be able to take an existing engagement, redesign it around outcomes, transfer, and leverage, and rewrite success so the client becomes stronger rather than merely serviced.

That is the sixth playbook move in the manuscript.

The next question is what kind of firm, talent mix, and operating model can actually deliver this consistently.

Chapter 11 — Talent and Operating Model

The next question is whether the firm itself is built to do any of this consistently.

The New Model Needs a Different Kind of Firm

The old delivery economy could survive a surprising amount of internal incoherence.

If enough capable people were available, if enough work could be staffed, if the client kept buying effort, and if a few senior operators held the whole thing together, the business could still function.

The multiplier path is less forgiving.

It depends more on judgment, clearer methods, stronger apprenticeship, better

knowledge transfer, and internal systems that help learning compound across engagements instead of disappearing into the next project kickoff.

That means the new model needs a different kind of operating discipline, not
just a more modern sales pitch.

Who Matters More After AI

Some roles become more important as routine effort compresses.

The first group is the one this book has already named: the people who carry
context, continuity, trust, and cross-functional translation. These are still the people you cannot afford to cut casually.

The second group is what you might call multiplier roles. These are people who
get clearer, not weaker, when AI reduces drag. They can frame better problems,
spot patterns faster, redesign workflows, teach others, or turn messy client situations into usable next moves.

The third group is apprenticeship nodes. These are the people and structures through which less experienced consultants learn judgment, not just tasks. If
the old junior-heavy pyramid weakens, firms have to become more intentional about how they develop people into stronger operators. Otherwise they protect
today's experts and accidentally run out of tomorrow's.

This is where the Do Not Cut List becomes more concrete. Protect the people
who carry continuity. Protect the people who multiply others. Protect the parts
of the system where judgment gets taught.

Apprenticeship Cannot Be Left to the Old Pyramid

One of the quietest risks in this transition is developmental.

The old model had flaws, but it did create repetition at scale. Junior people got exposure through volume. They learned the language, the patterns, the edge
cases, and the rhythms of client work by being close to lots of tasks.

If AI compresses much of that repetitive work, firms lose more than cost. They
lose part of the old learning path.

That does not mean apprenticeship disappears. It means apprenticeship has to be redesigned.

Some learning will now come less from repetition and more from guided method, joint problem solving, visible reasoning, better review loops, and more intentional transfer from experienced operators to newer ones. In other words, firms need to teach more of what they used to let people absorb indirectly.

That is more demanding. It is also more honest.

I saw a small version of that culture shift happen on one of my own teams when ChatGPT first arrived. I was pushing it hard, teaching people how to use it, doing lunch-and-learns, and talking about it enough that the team gave me the nickname GPLee. Then annual review season came around and I was blown away by the quality of Don and James's self-reviews. I asked how they had written them. They looked at me like the answer was obvious: "Duh, Lee. We used ChatGPT like you trained us to do."

That moment mattered because it showed the difference between one person using AI well and a team internalizing a new way of working. Apprenticeship in this era is not only about tasks. It is also about modeling curiosity, judgment, and tool use well enough that people keep doing it when you are not in the room.

Method Has to Become Visible

This is where many firms still rely too heavily on heroics.

A few strong people know how to diagnose the client well, frame the work well, steer the team well, and recover when things wobble. The firm calls them leaders, experts, or rainmakers. The problem is not that these people exist. The problem is that too much of the firm's real method still lives only inside them.

That is not a stable operating model for the next era.

The firm has to make more of its method visible, teachable, and reusable. Not every judgment can be turned into a checklist, and it should not be. But more of the scaffolding can be shared: how engagements are framed, how knowledge is captured, how quality is reviewed, how AI is used, how deliverables are shaped, and how learning gets carried from one project into the next.

I have spent a lot of my career building centers of excellence, solving recurring delivery problems, building tools around them, and enabling other people through documented methods. When I started developing software with AI and AI agents, I kept running into a new version of that same pattern. Context drift. Session loss. Uneven quality. False signals of completion. Tooling that was powerful but not yet dependable on its own.

Every meaningful piece of AgentFlow came from one of those scars. Product definition files exist because agents drifted scope on me. Design files exist because an agent made architectural decisions too silently. My testing rules exist because an agent once told me the work succeeded when the app did not really work. So I did what I had spent a career doing. I turned repeated problems into a more formal method. I called it AgentFlow.

That is the part that matters here. AgentFlow was not really a brand-new instinct. It was the same COE pattern applied personally: the methodology was the COE, the books became the curriculum, and the skills acted as governance. The point was not to replace judgment. It was to create a system that let judgment compound.

One of the strangest proofs of that for me was realizing I had written a book I could then read and learn from. The process was disciplined enough that it could help me come up to speed on the subject while I was still building the system around it. That is not an argument for pretending expertise does not matter. It is an argument for making method strong enough that learning compounds instead of resetting every time.

That pattern matters because it shows what a repeatable operating model really

is. It is not a speech about excellence. It is a way of turning good work into method, method into teaching, and teaching into better work the next time.

That is the operating-model lesson here. A firm that wants the multiplier path
has to build ways for good work to compound instead of starting from zero every
time.

Internal Discipline Is Now a Revenue Issue

Some leaders still treat internal operating discipline as overhead.

In the old model, they could sometimes get away with that. In the new model, it
becomes a revenue issue.

If methods are invisible, training is ad hoc, AI use is uneven, quality review is inconsistent, and learning disappears at the end of each engagement, the firm will struggle to deliver the new portfolio with confidence. The work will
depend too heavily on a few exceptional people. Growth will stall because the
model cannot repeat itself.

This is why internal systems matter more now. Reusable methods give people a
stronger starting point than memory and improvisation. Training loops turn what
used to spread by osmosis into something teachable. Stronger review practices
make quality less dependent on catching the right expert at the right time. Deliberate knowledge capture keeps the end of one engagement from becoming the
beginning of the next one all over again. Clear guidance on where AI belongs and where judgment still has to lead keeps leverage from turning into drift.

Those things are not back-office decorations. They are part of the product.

What a Smaller Firm Should Build First

Mid-market and boutique firms do not need a giant internal transformation program to make this real.

They do need a few things to become more deliberate:

- who the firm is protecting
- how newer people learn
- what methods should be visible

- how quality gets reviewed
- where knowledge goes after the work is done

Start there.

The first move is not to build an internal empire. It is to stop pretending the
new model can run on informal carry alone.

That is the same lesson as the service-portfolio chapters. One adjacent move beats a grand reinvention. One useful internal system is better than a speech about culture.

What This Means for a Firm of One

The same pattern holds for independent consulting.

A solo operator does not have a formal org chart, but they still have an operating model. They still have habits, methods, training loops, review practices, and ways of capturing or losing learning.

If all of that stays informal, the independent remains dependent on memory, energy, and personal heroics. That makes growth brittle.

This is one reason personal systems matter so much. The individual version of a
center of excellence might look like methodology, templates, review routines, AI workflows, and a deliberate way of capturing what worked so the next engagement starts higher than the last one.

That is not bureaucracy. It is self-compounding practice.

Build a Firm That Can Repeat the Better Model

The old habit is to think of talent, method, and internal discipline as support functions around the real business.

The better habit is to see them as the machinery that makes the new business
possible.

By this point, a consulting leader should be able to identify the talent archetypes that matter most after AI, redesign apprenticeship more intentionally, and name the internal systems and governance needed to make the
new model stick.

That is the seventh playbook move in the manuscript.

The next chapter turns outward again, because even a better firm still has to
open the right conversation with the client.

Chapter 12 — The Client Conversation

By the time a consulting firm has done the internal work this book describes, one question still determines whether any of it matters:

Can you open the right conversation with the client?

This is where good internal thinking often dies. The firm sees the trap more clearly. It knows what to protect. It has a better sense of what to sell and how to deliver it. Then it walks into the client conversation and defaults back
to the safest familiar language: efficiency, speed, automation, headcount, cost.

That language is not always wrong. It is often too small.

Start by Respecting the Efficiency Instinct

The fastest way to lose credibility in an AI conversation is to act as if cost pressure is fake.

It is not fake. Clients do want proof. They do want savings. They do want to
know whether the same work can now be done faster, with fewer delays, less manual effort, and less unnecessary labor.

Respecting that instinct matters because it is usually where the conversation starts.

The mistake is not beginning there. The mistake is letting the conversation end
there.

The Better Conversation Begins with a Harder Question

Once the efficiency case is on the table, the consulting leader has to widen the frame.

Not by sounding vague or visionary, but by asking a harder question:

If this initiative creates real capacity, where should that capacity go?

That question changes the meeting. It moves the client from cost accounting to

strategy. It forces a choice. It also exposes whether the sponsor is trying to optimize the present or build a stronger future.

This is the moment many firms miss. They answer the client's first question and
never ask the second one.

Use the Hidden Loss to Reframe the Stakes

The next move is to make the invisible visible.

What part of this work is genuinely getting cheaper? That is a fair question. What would be lost if all of those gains were harvested only as labor reduction? That is the one many leaders have not been asked with enough force.

This is where the ideas from earlier chapters become client language:

- which people are carrying continuity you cannot see on a spreadsheet?
- what judgment would walk out the door with an aggressive reduction move?
- what future capacity would be cut before it had a chance to matter?

That does not make the conversation sentimental. It makes it complete.

The Lewis and Clark Move

I saw this most clearly when I put forward what I called the Lewis and Clark
expedition proposal.

The organization already had large-scale AI adoption plans in motion. The real
constraint was more familiar: they had not simply opened access to the latest
tools for everyone, because large companies move at the pace of security. So the right first conversation was not "let's improve one process and remove cost." It was "how do we safely enhance exceptional employees so they become
meaningfully more productive, and then learn from that?"

That was the heart of the proposal. Start with five exceptional people. Give them unlocked-down Macs and subscriptions to the best available models. Let me
mentor them the way I mentor myself when I adopt new tools: not as passive users, but as high performers learning how to work with real leverage. Build in
security mitigations from the start. Make the first phase small enough to control, serious enough to matter, and explicit enough to learn from.

This was not a cost program. It was an employee amplification program. The goal
was not to remove labor from a process or make one workflow cheaper. It was to
see what happened when capable people were given real leverage, and whether
that capability could be developed, transferred, and scaled. That is a
different question. It is also the more important one.

The Lewis and Clark metaphor made that easier to hear. First footprints in the
snow. Then a wagon trail. Then, only after the lessons were clear, a highway
system across the enterprise. The first phase was never meant to be the whole
rollout. It was meant to prove whether enhanced employees, real tools, and
guided practice would create a level of productivity worth scaling. The wagon
trail phase would widen access with more control and safety. The highway phase
would bring mature enterprise governance, broader rollout, and clearer success
criteria.

That is the pattern worth copying.

Clients do not always need a grand transformation speech. Often they need a
credible first expedition: bounded enough to approve, structured enough to
learn from, security-aware enough to survive real enterprise scrutiny, and
strategic enough to point beyond short-term efficiency. The point is not merely
to automate a task. It is to discover what happens when capable people are
given real leverage under conditions the organization can trust.

The Conversation Moves That Open Better Work

In practice, the sequence is straightforward.

First, validate the efficiency instinct. Second, expose the hidden loss in
reduction-only thinking. Third, redirect the conversation to growth allocation.
Fourth, define what must be protected. Fifth, propose a better first
engagement.

That sequence works because it respects the sponsor's actual pressure while
still leading them somewhere better.

In a real meeting, the goal is not to recite these lines. It is to listen for

where the sponsor is stuck and then use the next move that widens the frame
without losing credibility. It might sound like this:

"Yes, some of this work should get cheaper. The question is what you want to do
with the capacity that creates."

"If you take every gain as headcount reduction, what knowledge or continuity goes with it?"

"Which people or functions are carrying more than the org chart shows?"

"If this first phase works, what should be true beyond savings?"

"Would you rather run a controlled first expedition that proves the right things, or optimize the current work faster and still not know where the capacity should go?"

The pattern to listen for is simple. Start where the sponsor already agrees. Then move toward what they are not yet seeing. That is not a script. It is a set of moves.

The Objections You Should Expect

The first objection is usually some version of: "We need hard ROI first."

That is fair. The right answer is not to dismiss ROI. It is to broaden what counts as proof. Cost savings are one kind of proof. So are stronger internal capability, protected continuity, faster decisions, better workflows, and a clearer destination for new capacity. When this move works, the sponsor stops
treating ROI as only a labor question and starts treating it as a capability question too.

The second objection is: "This sounds softer than efficiency."

It is important to reject that framing. Growth allocation is not softer. It is more demanding. It requires clearer ownership, better tradeoffs, stronger design, and more discipline than a simple reduction move. When this answer lands, the conversation shifts from sentiment to standards.

The third objection is: "We are not ready for a big transformation."

That is where the Lewis and Clark pattern helps. Lower the scope without lowering the seriousness. Propose a bounded first phase with clear outputs, learning transfer, and a defined question it is supposed to answer.

What This Means for a Firm of One

Independent consultants need this chapter too.

A solo operator will often feel pressure to make the AI conversation sound as
cheap, fast, and low-friction as possible. That can win a meeting. It can also
trap the offer in the wrong category from the beginning.

The same conversational move still matters. Yes, some work is faster now. But
what should that new capacity be used for, and what should not be lost in the
process?

That is how an independent avoids becoming "the cheaper way to get the same
work done" and becomes the person who can frame what the better work should be.

Say the Harder Thing Early Enough to Matter

The old habit is to answer the client's first question and stop there.

The better habit is to answer it and then move quickly to the more strategic one.

That is how the consulting firm earns the right to a better engagement.

By this point, a consulting leader should be able to reframe an AI conversation
from reduction to growth, challenge reduction-only logic without sounding naive, and open a better first engagement with a skeptical sponsor.

That is the eighth playbook move in the manuscript.

The final chapter now has one job left: show what the winning post-delivery firm looks like when all of this starts to hold together.

Chapter 13 — The New Consulting Firm

At some point, a book like this has to stop breaking the old model apart and
show the reader what the better one looks like when it starts to hold together.

That is the job of this chapter.

One point is worth saying plainly before I try to describe that firm. I have never run a consulting firm. I have spent more than three decades inside

them,
building practices, leading delivery, selling work, shaping methods, building centers of excellence, and being the kind of person this book keeps arguing you
should not cut too quickly. That is not managing-partner authority. It is witnessed consequence. For this argument, I think that matters.

The new consulting firm is not a fantasy firm. It is not frictionless. It is not perfectly automated. It is not a pure strategy boutique floating above the
mess of implementation. It is a real firm that has learned how to live in a world where software delivery is cheaper, generic execution is easier to reprice, and clients can see more of the production machinery than they used
to.

What makes it different is not that it has escaped those forces. It has built around them more intelligently.

It Sells a Different Kind of Value

The new consulting firm still delivers work. It still solves hard problems. It still needs people who can execute.

But it no longer treats visible labor as the center of gravity.

Its value sits more clearly in judgment, redesign, capability building, knowledge preservation, operating method, and the ability to turn AI-enabled
productivity into stronger client outcomes. It is easier for the firm to say what clients are buying beyond "a good team doing a lot of work."

That shift changes everything downstream.

The portfolio looks different. The engagement model looks different. The talent
system looks different. The client conversation sounds different. The internal operating model grows more deliberate because the firm can no longer rely on
the old pyramid to carry so much of the learning and economics by default.

It Uses AI as a Force Multiplier, Not a Shrink Ray

This may be the clearest difference of all.

The winning firm is not anti-efficiency. It uses AI aggressively where AI creates real leverage. It removes drag, improves first passes, accelerates synthesis, strengthens options, and helps strong people operate with much

more
range.

What it does not do is mistake those gains for a reason to hollow itself out.

The firm has learned the difference between a force multiplier and a shrink ray.

That sounds simple. It is not. It requires leadership restraint at exactly the moment short-term savings are easiest to celebrate. It requires the firm to protect continuity, preserve learning paths, and reinvest part of the gain into
future capability and future offers rather than taking every visible win as a reason to get smaller.

That is why the winning firm is usually more deliberate than dramatic.

It Is Built to Compound

The older delivery economy tolerated too much reinvention. Teams started over.
Methods stayed half-hidden. Learning got trapped inside strong individuals. Good work was real, but too much of it did not compound.

The new consulting firm is built to compound.

It captures more of its method. It teaches more of its reasoning. It makes more
of its quality visible and reviewable. It gets better at transferring knowledge to clients without giving away the whole business. It improves its own practice
through reuse, governance, and explicit learning loops instead of relying on informal carry from one heroic operator to the next.

That is what makes the model scalable without making it generic.

This matters beyond consulting. Some of the clearest proof in my own life came
when AI helped me develop and express ideas I had carried for years with more
range and rigor than I could manage alone. But the lesson is not that AI lets
people safely work beyond their real expertise. Serious work still requires real judgment and the ability to supervise what the tool is producing. The more
defensible lesson is narrower and more human: leverage can help a grounded person express, test, and extend what they already genuinely carry. That is part of what this book means by a more human firm. The promise is not only that

work gets faster. It is that capability that used to stay trapped inside a person, a role, or a routine can finally find room to become visible.

It Develops People Differently

The firm still needs talented individuals. It also understands that talent alone is not the model.

The roles that matter most are clearer now. Knowledge Carriers matter more.
Multiplier roles matter more. Apprenticeship nodes matter more. The firm knows
who it cannot afford to cut lightly, who it should be amplifying, and where judgment gets taught.

Because the old learning pyramid is weaker, the winning firm becomes more intentional about development. It teaches more of what used to be absorbed by
proximity and repetition. It creates stronger review loops. It treats method as
part of the product and apprenticeship as part of the business model.

That does not make the firm bureaucratic. It makes it less wasteful.

It Looks Different at Different Scales

The model is not identical in every firm.

In a large consultancy, the new model may show up first as a sharper split between differentiated advisory, capability-building work, and delivery work that is still necessary but no longer allowed to define the firm's future. The large firm still benefits from scale, brand, client access, and cross-capability breadth. Its challenge is to keep those advantages from becoming excuses for
keeping too much of the old portfolio untouched.

In a boutique or mid-market firm, the model usually shows up as sequence. One
offer gets redesigned. One adjacent offer gets built. One internal method gets
made visible. One better client conversation opens a better engagement. The
firm moves by a series of sharper decisions rather than by grand transformation theater.

In a firm of one, the model is more personal but not less real. The independent
stops selling only time and visible effort. They build methods, package

judgment, show a stronger point of view, and use AI to increase the level of problem they can help clients solve rather than only the speed at which they
solve the old one.

The scale changes. The logic does not.

It Earns Better Dependence

This is one of the most important shifts in the book.

The old model often depended on client dependence of the wrong kind: we keep
doing this work because the client still needs us to repeat it.

The better model earns a better kind of dependence.

The client depends on the firm because the firm helps it think better, redesign
better, preserve what matters, build stronger capability, and move faster on the decisions that shape the business. That is a more defensible relationship because it is harder to replace with cheaper labor or a faster tool.

The firm still creates follow-on work. It just creates it at a higher level.

What Winning Actually Looks Like

Winning in this era does not mean becoming the most automated firm in the
market.

It means becoming one of the firms that knows what automation is for.

Picture the firm three years from now. A partner meeting is underway, but the
conversation sounds different from the old model. No one is asking only how
many hours can be sold or which team can be staffed more cheaply. They are
looking at which methods are compounding, which clients are moving into deeper
work, which people have become more valuable with leverage, where new capacity
went, and which follow-on offers are growing because the earlier engagements left the client stronger. In another room, a newer consultant is learning from
visible methods and review loops that did not exist before. In a client meeting
later that day, the firm is not selling speed alone. It is selling a better

path through complexity. That is what the model looks like when it becomes real.

It means the firm can explain what it protects, what it amplifies, where new capacity goes, what work it is leaving behind, what new work it is building, how it delivers that work, how it develops people, and how it opens the client
conversation that makes the whole model real.

That is a very different standard from "we are using AI now."

It is also a much harder one.

The Firm This Book Has Been Pointing Toward

The book has been building toward a firm that is more human, not less disciplined. More productive, not more hollow. More explicit about method, not
less creative. More serious about growth, not less serious about economics.

That firm will still feel pressure. It will still make mistakes. It will still need to adapt as the tools improve and the market changes again.

But it will not be trapped in the old logic.

It will know that cheap code is not the same thing as cheap judgment. It will
know that AI value is not fully captured when a process gets cheaper. It will
know that the right people become more important when leverage expands, not
less. And it will know that the future belongs to firms that can turn capability into compounding business value rather than only into lower labor cost.

That is the firm worth building.