

AI VALUE MANAGEMENT

The AI Value

*How Management Turns AI Adoption
Into Measurable Business Value*

L A Y E F A O K I Y

Gap

The AI Value Gap

*How Management Turns AI Adoption Into Measurable
Business Value*

LAYEFA OKIY



Copyright © 2026 Layefa Okiy. All rights reserved.

No part of this book may be reproduced or transmitted without permission, except for brief quotations used in reviews or scholarly discussion.

Acknowledgements

This book was shaped by the researchers, operators, managers, public institutions, and practitioners whose work made its evidence base possible. The author also thanks the colleagues, readers, and critics who challenged the argument and helped turn a broad concern about AI adoption into a practical management discipline.

Contents

Introduction · The Buildout Is Real. The Value Gap Is Too.	9
PART I · THE AI VALUE GAP	17
Chapter 1 · Adoption Is Not Value	18
Chapter 2 · Capability Is Not Strategy	26
Chapter 3 · The Tool-First Trap	33
Chapter 4 · AI Scales the Organization You Already Have	41
PART II · THE MANAGEMENT SYSTEM FOR AI VALUE	53
Chapter 5 · AI Value Management and the Five Conditions	54
Chapter 6 · Define the Value Case Before the Use Case	70
Chapter 7 · From Output to Outcome	83
PART III · DESIGNING AI INTO THE WORK	99
Chapter 8 · Where AI Belongs in the Workflow	100
Chapter 9 · Which Model Belongs in the Work?	109
Chapter 10 · Decision Rights, Human Judgment, and Escalation	117
Chapter 11 · Governance as Operating Infrastructure	128
PART IV · FROM READINESS TO RUNTIME	137
Chapter 12 · Ready for What, Exactly?	138
Chapter 13 · Pilots That Produce Decisions	147
Chapter 14 · Keep Value and Permission Current	157
PART V · REPEATING THE DISCIPLINE	167
Chapter 15 · Managing the AI Portfolio	168
Chapter 16 · Building the AI Value Organization	178
Conclusion · Close the Gap	187
Appendix A · How AI Value Management Relates to Existing AI Governance and Management Systems	192
Glossary	194
Notes	197
Index	204
About the Author	206

Figures and Exhibits

Figure 1.1	The Adoption-to-Value Ladder	22
Exhibit 2.1	The Strategic AI Test	29
Figure 3.1	The Tool-First Trap	36
Figure 4.1	What AI Amplifies	44
Figure 5.0	AI Value Management at a Glance	59
Figure 5.1	The Five Conditions of AI Value	60
Figure 6.1	The AI Value Case	72
Figure 7.1	The Output-to-Outcome Chain	87
Exhibit 7.1	What 220 public AI claims could show	86
Figure 8.1	Where AI Belongs in the Workflow	102
Figure 9.1	Two chains connect capability to value	113
Figure 10.1	Decision Rights, Review, and Escalation	119
Figure 11.1	Governance as Operating Infrastructure	131
Figure 12.1	Readiness for the Next Commitment	141
Figure 13.1	The Managed Pilot Contract	150
Figure 14.1	The Runtime Assurance Loop	159
Exhibit 15.1	Boundaries improve inspectability	172
Figure 15.1	The AI Portfolio Allocation Map	171
Exhibit 15.2	Portfolio Trade-off Record	175
Figure 16.1	The AI Value Organization	181
Exhibit 16.1	The First Ninety Days	184

How to Use This Book

The book is written to be read in order. Part I removes four illusions. Part II builds the management system. Part III designs AI into the work. Part IV moves from readiness to runtime. Part V makes the discipline repeatable across the enterprise.

WHERE TO START

Chief executives and boards	Introduction · Chapters 1, 2, 5, 15, 16 · Conclusion
Finance leaders	Chapters 1, 6, 7, 9, 15
Operations, product, and technology leaders	Chapters 3, 4, 8, 9, 10, 13
Risk, legal, and assurance leaders	Chapters 10, 11, 12, 14

READING AIDS

Chapter at a glance	Each chapter opens with the question it answers, its five steps, and its decision rule. From Chapter 5, the opener also marks which of the Five Conditions the chapter develops.
Decision and Boundary	The rule to apply, and the line the evidence must not cross.
In practice	How to put the argument to work, and the warning signs to watch for.
Case sidebars	Company cases, read for their management lesson.
Exhibits	Figures and data, each with its source.
Chapter summary	Key takeaways and questions for the leadership team close Chapters 1 to 16.

The Glossary defines the book’s terms. The Index locates cases and concepts. The companion toolkit carries the full operating instruments.

The Book at a Glance

Each chapter ends in one decision rule. Read together, the rules are the discipline.

The Buildout Is Real. The Value Gap Is Too.		Evidence should authorize the next commitment – not the strongest story the organization can tell.
PART I • THE AI VALUE GAP		
1	Adoption Is Not Value	Do not ask adoption evidence to prove value. Ask what it can authorize next.
2	Capability Is Not Strategy	Capability creates options. Strategy decides which options deserve commitment.
3	The Tool-First Trap	Make the operating decisions before product momentum makes them implicitly.
4	AI Scales the Organization You Already Have	Examine the organization that will receive the capability before asking what the AI can do.
PART II • THE MANAGEMENT SYSTEM FOR AI VALUE		
5	AI Value Management and the Five Conditions	The burden should match the decision being authorized.
6	Define the Value Case Before the Use Case	Approve the uncertainty worth testing – not the technology the organization already wants to scale.
7	From Output to Outcome	Evidence at one level authorizes only the next justified claim.
PART III • DESIGNING AI INTO THE WORK		
8	Where AI Belongs in the Workflow	A placement is credible only when both upstream and downstream are designed.
9	Which Model Belongs in the Work?	Choose the configuration after the work is known – and let the decision expire.
10	Decision Rights, Human Judgment, and Escalation	A review requirement with no capacity model is an unfunded control.
11	Governance as Operating Infrastructure	Governance creates speed when the safe path is also the easy path.
PART IV • FROM READINESS TO RUNTIME		
12	Ready for What, Exactly?	The largest justified commitment follows the weakest unresolved critical gate.
13	Pilots That Produce Decisions	A pilot succeeds when it changes what management is justified in doing.
14	Keep Value and Permission Current	Continue, restrict, reauthorize, and exit are all management decisions.
PART V • REPEATING THE DISCIPLINE		
15	Managing the AI Portfolio	A portfolio is managed only when the review changes a commitment.
16	Building the AI Value Organization	Make accountable AI decisions repeatable, and revise them when conditions change.
	Close the Gap	AI capability can be purchased. The management capacity to turn it into value must be built.

Intro

INTRODUCTION

The Buildout Is Real. The Value Gap Is Too.

THE QUESTION

Why does AI value lag AI adoption, and what closes the gap?

THE ARGUMENT IN FIVE MOVES

- | | | | | |
|----------------------|------------------------------|---|----------------------------|-------------------------------|
| 01 | 02 | 03 | 04 | 05 |
| The buildout is real | The evidence locates the gap | The hard part moved into the organization | No function owns the chain | A management system closes it |

DECISION RULE

Evidence should authorize the next commitment – not the strongest story the organization can tell.

IN THIS CHAPTER

The contradiction inside the organization	10	The difficult part has moved into the organization	13
What the current evidence establishes	11	The missing management system	14
The value may be delayed – but the path still matters	12	The journey of the book	14

THE WORLD IS committing capital to artificial intelligence faster than most organizations can explain what they are receiving in return.

The commitment is visible in concrete, steel, silicon, electricity, fibre, software, and debt. BlackRock Investment Institute's June 2026 outlook describes an accelerating AI buildout constrained by scarce power, memory, chips, data centres, labour, and capital.¹ Stanford's 2026 AI Index reports that global corporate AI investment reached \$581.7 billion in 2025, more than double the previous year.² These measures describe investment, not return. Their importance is that AI has moved beyond a software-feature cycle. It is becoming an industrial and organizational project.

The pressure inside companies is rising with the spending. BCG's 2026 AI Radar surveyed 2,360 executives, including 640 chief executives.³ Respondents expected planned AI investment to rise from roughly 0.8 per cent of revenue in 2025 to about 1.7 per cent in 2026. Ninety-four per cent said their organizations would continue investing even if current initiatives did not produce the desired financial impact over the following twelve months. Seventy-two per cent of CEOs described themselves as the main AI decision-maker, and half believed their job stability depended on getting the strategy right.

Those figures reveal more than optimism. They reveal an incentive system. The executive responsible for AI is being asked to accelerate before the organization has finished building the workflow, measurement, ownership, data, and control systems that make value visible. Moving slowly creates strategic and reputational risk. Moving quickly without those conditions creates operational and financial risk.

Governments are adding another layer of pressure. Canada's 2026 national strategy aims to raise business AI adoption from 12 per cent to 60 per cent by 2034 and to support as many as 250,000 jobs through adoption by 2031.⁴ The public direction is unmistakable: organizations are being encouraged to adopt, commercialize, and scale.

The buildout is real.

The value gap is real too.

AI adoption is everywhere. AI value is not.

AI already improves products, tasks, workflows, and firms, and many organizations are making money from it. The gap is that access, use, and spending have advanced faster than the management systems required to turn capability into repeatable outcomes.

The contradiction inside the organization

Generative AI made useful experimentation unusually easy. An employee could open a browser, enter a task, and receive a plausible result within minutes. A business unit could

test a vendor without rebuilding its technology estate. A software provider could add AI to a product customers already used.

That accessibility produced genuine discovery. It also collapsed distinctions management depends on.

Access began to look like adoption. Adoption began to look like deployment. Deployment began to look like transformation. A faster draft began to look like productivity. A collection of pilots began to look like a strategy. A rising usage curve began to look like value.

The labels are not harmless. Each one should support a different decision.

A company can distribute an assistant to ten thousand employees, observe that thousands use it every week, and receive survey estimates of time saved. That is meaningful evidence of availability, behaviour, and perceived task benefit. It does not yet show whether output quality improved, whether review work moved to other employees, whether confidential information entered unapproved systems, whether saved capacity was productively absorbed, or whether cost, revenue, service, risk, or strategic position changed.

Every number on the board slide may be accurate. The conclusion can still be wrong.

The pressure to tell the optimistic story is structural. Vendors can report adoption quickly. Innovation teams can count pilots. Technology teams can count licences. Employees can report convenience. Executives can show movement. Outcome evidence arrives later, crosses functions, and creates harder questions about attribution, cost, customer consequence, role design, and whether the investment thesis was right.

Activity normally has an owner. Value often falls between owners.

What the current evidence establishes

The 2026 evidence points to a management problem.

Across those different measures, the pattern is consistent. PwC's 29th Global CEO Survey found that 12 per cent of respondents reported both lower costs and additional revenue from AI, while 56 per cent reported no significant financial benefit to date.⁵ Deloitte distinguished surface-level use from process redesign and deeper transformation.⁶ McKinsey's measurement work connected technical health and adoption to operational performance, strategic outcomes, full cost, attribution, and financial impact.⁷ Each study measures something different – use, deployment, financial benefit, EBIT impact – so the figures are read side by side, not added together.

The pattern is consistent: spending and access are moving faster than workflow redesign, ownership, measurement, and organizational learning.

My own 2026 research found the same pattern in public evidence: AI is often visible in live work long before the public record shows what changed, what the work fully cost, or how much of the reported result can reasonably be connected to the AI itself.

Scalebotics Research, under my direction, screened 200 companies and 300 public AI claims.⁸ Of the 220 claims with enough public material to profile, 193 showed good evidence that the AI was running in real work. Only 47 provided a believable before-and-after or comparison point,

■ EVIDENCE

Of 220 profiled public claims, 193 showed AI in live work. Only 16 showed strong evidence tying the result to the AI. Operation is easy to show; attribution is not.

36 gave meaningful evidence about cost, savings, revenue, or return, and 16 showed strong evidence connecting the reported result to the AI itself.

Across those 220 public claims, activity was easier to demonstrate than consequence. The evidence available should therefore authorize the next commitment, not the strongest story the organization can tell about the initiative.

The counterevidence matters just as much. Controlled experiments and field studies show that AI can reduce task time, improve output quality, and increase productivity in well-defined settings.⁹ Those findings do not weaken the value-gap argument. They locate it.

If AI could not improve useful tasks, organizations would stop buying it. The management problem exists because the technology can produce visible and sometimes impressive task gains while the route from those gains to enterprise performance remains incomplete.

A writing assistant can help one employee finish faster. Management still has to decide whether the organization needs more writing, better writing, lower cost, faster response, stronger compliance, or a different allocation of human effort. A support assistant can help agents resolve more cases per hour. Management still has to consider demand, staffing, service quality, escalation, customer trust, integration cost, labour effects, and what happens when the assistant is wrong.

The model creates capability. The organization determines its result.

The value may be delayed – but the path still matters

The strongest counterargument deserves respect: perhaps much of the value is real but delayed, localized, intangible, or badly measured.

General-purpose technologies often require complementary investment before their largest benefits appear.¹⁰ Electricity did not transform factories merely because motors became available. Organizations redesigned production. Computers did not create their

greatest gains by replacing typewriters one for one. Firms changed software, data, training, products, coordination, and work.

Artificial intelligence may follow part of that pattern. Some investments create learning or option value before they produce earnings. Some benefits appear in quality, resilience, customer experience, workforce capability, or strategic position. Conventional accounting may record the cost before it recognizes the capability being built.

But a lag is an explanation, not a blank cheque.

If value is expected later, management still needs milestones, owners, evidence, and a theory of how the investment matures. If the benefit is capacity, leaders must say how capacity will be used. If it is learning, they must name the uncertainty being reduced and the decision that learning will support. If it is strategic, they must identify the option, asset, position, or capability being created.

“AI is a long-term investment” can be a rational position. It can also become a refuge from measurement. The difference is whether management can describe the path.

■ IN PRACTICE

When a sponsor says value will arrive later, ask for three things: the milestone, the owner, and the date on which an absent result counts against the investment.

The difficult part has moved into the organization

The most visible AI investment is usually the model, licence, or software interface. The decisive investments are often less visible.

They include understanding the workflow well enough to know where AI belongs; connecting the system to authoritative data; defining what outcome matters; training people to use the capability and recognize its limits; assigning authority to approve, override, suspend, and review; designing what happens when the ordinary path fails; measuring cost and consequence after launch; and updating the system when the model, vendor, data, policy, market, or surrounding process changes.

These are not support activities around the real AI work. They are the real work of converting capability into value.

The hard part has moved upstream into problem definition, context, evidence, and authority – and downstream into review, workflow redesign, implementation, measurement, and accountability.

This is easy to flatten into the instruction to focus on people, process, and technology. That phrase is too broad to guide a consequential decision. Which people? Which process? Which outcome? Which owner? Which data? Which boundary? Which evidence? Which failure can be tolerated? Which action requires permission? Which capacity will be released? Which stakeholder receives the benefit? Which stakeholder carries the downside? What would make the organization stop?

The answers differ by use. A private assistant helping one employee summarize internal notes does not require the same operating design as an agent changing a customer account, a model recommending credit, a system screening applicants, or an application generating clinical documentation.

AI adoption hides those differences. Management must make them visible.

The missing management system

Most organizations already possess many of the relevant functions. Strategy can explain where the business intends to compete. Technology can make the system run. Operations can describe the workflow. Finance can challenge the economics. Risk, legal, privacy, security, and compliance can define boundaries. Human resources can support role change. Domain specialists can judge whether an output makes sense.

Each function owns a legitimate part of the answer. No function necessarily owns the chain.

That is how an organization can have an AI strategy, a governance policy, an inventory, a project office, approved vendors, and dozens of pilots while still being unable to answer a basic question: is this specific business use of AI creating worthwhile value, and what decision should the current evidence authorize?

Between strategy and outcome lies a management system. It chooses the problem. It states the value case before the tool creates momentum. It locates AI in the work. It defines what the system may do, recommend, or never do without approval. It assigns authority. It designs review, exception, escalation, and remedy. It judges readiness. It runs experiments to reduce named uncertainty. It measures the result. It keeps permission current after deployment. It allocates scarce capital and specialist capacity. It stops uses whose conditions no longer hold.

This book calls that discipline AI Value Management.

The name is not the contribution. The contribution is the connected system – and the claim that management must treat those decisions as one value-producing architecture rather than as separate specialist tasks.

■ KEY TERM

AI Value Management

The discipline of selecting, designing, authorizing, operating, proving, and adapting specific business uses of AI so that capability becomes worthwhile, measurable, and sustainable value.

The journey of the book

The book moves through five Parts.

First, it removes the illusions that make bad decisions look reasonable: adoption is not value, capability is not strategy, a tool is not an operating system, and AI does not enter a neutral organization.

Second, it defines the value system: what worthwhile value means, which conditions must hold, how to define the value case before the use case, and how evidence moves from output to outcome.

Third, it designs AI into real work: placement, model configuration, authority, review, escalation, and governance.

Fourth, it moves from readiness to runtime: the next justified commitment, pilots that produce decisions, and keeping value and permission current. Fifth, it makes the discipline repeatable: portfolio allocation and the organization required to build, operate, learn, and stop coherently across many uses.

The purpose is not to create another committee between the business and the technology. It is to create clearer decisions – fast where the work is reversible and bounded, more demanding where rights, safety, customer obligation, financial exposure, architecture, workforce structure, or long-term dependence are at stake.

The next competitive advantage will not come from access alone. It will come from the organization's ability to decide where AI matters, put it into the work, remain in command of what it may do, measure what changes, and revise the decision as conditions change.

That future is not reserved for companies that began as AI firms. It can be built deliberately by organizations willing to manage the conditions around the capability.

The governing reader question is therefore not only, why are we not seeing value? It is: where should we use AI, what must be true for it to create worthwhile value, and what evidence should authorize the next commitment?

The buildout is real. The value gap is real too. Closing it is the work of management.



PART I

The AI Value Gap

Remove four illusions before building the management system: adoption is not value, capability is not strategy, a tool is not an operating system, and AI does not enter a neutral organization.

01	Adoption Is Not Value	18
02	Capability Is Not Strategy	26
03	The Tool-First Trap	33
04	AI Scales the Organization You Already Have	41

01

CHAPTER 1 • PART I

Adoption Is Not Value

THE QUESTION

What does each level of AI evidence actually prove?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Activity wins the reporting contest	Six levels, six different claims	Each level authorizes a different step	Task gain is not system value	Saved time is capacity, not cash

DECISION RULE

Do not ask adoption evidence to prove value. Ask what it can authorize next.

IN THIS CHAPTER

1.1	The category error	19	1.5	Task gain is not system value	23
1.2	Why activity wins the reporting contest	20	1.6	When saved time becomes value	23
1.3	The ladder is also a decision chain	21	1.7	Different levels, different claims	24
1.4	The evidence funnel is the point	22	1.8	What the evidence authorizes	25

IN FEBRUARY 2024, KLARNA offered what looked like one of the cleanest enterprise AI success stories in the market.¹

The Swedish financial-technology company said its new AI customer-service assistant had handled 2.3 million conversations during its first month – roughly two-thirds of all customer-service chats. According to the company, the assistant was doing work equivalent to 700 full-time agents, had reduced repeat inquiries by 25 per cent, and had cut average resolution time from eleven minutes to less than two. Klarna expected the system to improve profit by US\$40 million during the year.

Those are not vague claims about transformation. They are concrete operating measures: volume, labour equivalence, repeat contacts, resolution time, and projected profit.

Eighteen months later, the company's story had become more complicated. In September 2025, Klarna's chief executive told Reuters that the company had probably over-indexed on AI-led cost cutting and was course-correcting.² The company was hiring people again and shifting attention toward service quality, products, customer experience, and growth.

The tempting conclusion is that the original deployment failed. That conclusion is too easy.

The original metrics may have been accurate. The assistant may have handled the reported volume. Resolution time may have fallen. Repeat inquiries may have declined. Labour requirements may have changed. The strategic problem is that none of those measures, individually, could settle the full question of value.

A system can be efficient at the task it performs and still be incomplete as a business strategy.

Klarna's course correction reveals the distinction at the centre of this chapter. Access to AI, use of AI, adoption of AI, operational deployment, enterprise scale, and measurable business value are not interchangeable. Each describes a different claim. Each requires different evidence. Progress at one level does not prove progress at the next.

DECISION

Evidence should authorize the next justified commitment – not the strongest story the organization would like to tell.

1.1 The category error

Organizations rarely announce that they have confused activity with value. The confusion arrives through a sequence of numbers that are individually true.

A chief information officer reports that twenty thousand employees have access to an assistant. A product team reports that sixty per cent activated it. An innovation team reports eighty pilots. A department reports that users save an hour each week. A vendor reports millions of prompts. A board slide places the figures under the heading “AI value delivered.”

The slide moves too quickly.

Licences prove access. Active users prove usage. Repeated use may prove adoption. Pilots prove experimentation. A reported hour saved suggests a possible task benefit. None, alone, proves that the organization is materially better after the costs and consequences of that AI use are counted.

■ IN PRACTICE

Relabel the board slide. Place each metric under the level it proves – access, usage, adoption, deployment, scale, or value – and remove any heading that claims more.

This is not simply a measurement mistake. It is an incentive-compatible reporting system.

Access and usage measures appear quickly. They are easy to collect and compare. Vendors can produce them automatically. Innovation teams can influence them directly. Executives can present them before the organization has completed a full operating cycle.

Outcome measures are slower and politically harder. To measure cost, someone must decide which costs belong to the AI use and whether saved capacity was actually removed or avoided. To measure quality, the organization needs a standard independent of the people using the tool. To measure customer impact, it may have to connect data across departments. To measure risk, it must record incidents and near misses that teams may prefer not to surface.

Activity has an owner. Value often falls between owners.

Technology owns the platform. Operations owns the workflow. Finance validates the economics. Risk defines boundaries. Human resources manages role changes. A business-unit leader owns the customer result. When those decisions are not connected, every function can complete its part while the overall benefit remains unowned.

The organization does not lack data. It lacks a value architecture.

1.2 Why activity wins the reporting contest

Activity metrics dominate early AI programmes for rational reasons. They are available sooner, more controllable, and politically convenient. A licence can be counted on the day it is issued; financial, customer, quality, and workforce outcomes usually require a baseline, a full operating cycle, evidence from several functions, and agreement about attribution.

A programme leader can distribute more licences, recruit pilot teams, increase training, and encourage use. The same leader may have limited authority over staffing, customer demand, operating budgets, role design, or whether another function absorbs the capacity created.

Activity also aligns with how the AI market is sold. Vendors can demonstrate what the model produces and provide dashboards showing engagement. They cannot, without access to the customer's economics and operations, prove how the organization converted that activity into value.

The result is a reporting system in which everyone can be successful before the organization is successful. The vendor achieves deployment. Technology achieves availability. Innovation achieves pilot volume. Employees achieve convenience. The executive achieves a credible story of movement. The missing question – what outcome changed enough to justify the system – belongs to no one.

This is not necessarily dishonesty. It is system design.

Organizations receive the measures their incentives, ownership, and reporting cadence make easiest to produce. When value appears later and crosses functional boundaries, management must deliberately create an owner and a measurement system for it. Otherwise, the programme naturally optimizes for the nearest visible proxy.

Proxies are not useless. They become dangerous when management forgets what they stand in for. A falling resolution time may be a valuable leading indicator. It becomes a weak value claim if service quality falls or difficult cases move into a more expensive channel. Weekly active use may show that employees find a tool helpful. It says little about whether a workflow, cost structure, customer outcome, or strategic position changed.

The discipline is not to eliminate proxy measures. It is to connect each proxy to a stated outcome and a decision.

1.3 The ladder is also a decision chain

The ladder matters because evidence at each level should authorize a different decision.

Evidence of access may authorize controlled experimentation. Evidence of usage may justify deeper research into the tasks attracting demand. Evidence of sustained adoption may justify workflow design and integration. Evidence of deployment may justify tests of reliability, economics, review burden, and operating consequence. Evidence of scale may support broader capital-allocation choices. Evidence of value may justify continued investment, expansion, redesign, or termination.

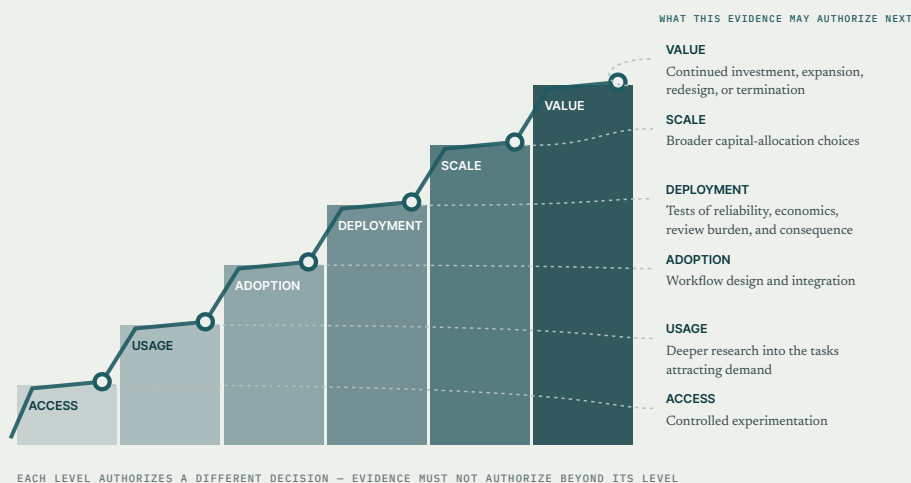
Problems begin when evidence is asked to authorize a decision beyond its level.

A high activation rate may justify more research. It does not necessarily justify enterprise rollout. A promising task experiment may justify a workflow pilot. It does not

justify restructuring the workforce. A successful pilot may justify investment in integration. It does not establish that the AI use will remain economic after support, monitoring, change, and exceptions are added.

FIGURE 1.1

The Adoption-to-Value Ladder



Each level proves something different and authorizes a different next decision.

SOURCE: AUTHOR.

This logic protects both speed and discipline. Low-consequence, reversible decisions can move on limited evidence. More consequential and difficult-to-reverse decisions require stronger proof. The organization does not wait for final ROI before learning, and it does not allow early enthusiasm to harden into expensive structure.

WATCH FOR

Evidence travelling upward: an activation rate cited to justify enterprise rollout, or a task experiment cited to justify restructuring the workforce.

1.4 The evidence funnel is the point

The public-evidence benchmark makes the distinction visible at scale. Of the 220 public AI claims in the 2026 Scalebotics benchmark, 193 showed good evidence of real use.³ Only 47 offered a believable comparison, 36 meaningful economics, and 16 strong evidence tying the result to the AI.

Evidence of operation is easier to produce and disclose than evidence of causal and economic consequence; the other 204 claims may describe real value that the public

record does not show. A system can be live, widely used, governed, and even associated with an improved outcome while the strongest value claim remains unauthorized.

That is the practical meaning of the ladder. Every level answers a different question. Evidence of access can justify learning about use. Evidence of use can justify testing workflow fit. Evidence of a bounded workflow effect can justify a stronger experiment. None of those steps entitles management to jump directly to an enterprise-value conclusion.

1.5 Task gain is not system value

A large field experiment involving Microsoft 365 Copilot makes the distinction visible.⁴ Researchers partnered with 66 large firms and randomly assigned access among 7,137 knowledge workers. More than 90 per cent of workers offered the tool used it at least once. After twelve weeks, weekly usage stabilized at just under 40 per cent.

The tool changed some work behaviour. Workers with access spent less time in Outlook and consolidated email work into fewer sessions. The researchers did not detect significant average changes in time spent in meetings or writing documents, and they could not directly observe productivity, performance evaluations, or work quality.

The evidence shows that some individual work patterns changed and some time was saved. It does not show that the firms redesigned workflows, increased enterprise output, reduced cost, or improved profit.

The same field evidence also shows why coordination matters. Individual workers can change email habits on their own; changing meetings, shared documents, role boundaries, handoffs, and team norms requires collective redesign. The larger organizational gain often depends on complementary changes that no individual user can make by simply becoming more proficient with the tool.

That is the adoption-to-value gap in miniature. The capability was real. The use was real. The task benefit may have been real. The organization still had to decide what outcome it wanted, what happened to the released capacity, and what broader claim the evidence justified.

1.6 When saved time becomes value

Hours saved is one of the most common AI value claims because time is intuitive and easy to monetize on a spreadsheet. The arithmetic often multiplies reported hours by salary cost and presents the result as ROI.

That calculation can estimate potential capacity. It is not automatically a realized financial benefit.

BOUNDARY

A reported hour saved estimates potential capacity. It is not automatically a realized financial benefit.

Employees may use the time to complete more work, improve quality, support customers, solve harder problems, learn, or innovate. They may finish earlier and reduce overtime or out-of-hours work. Demand may be fixed and staffing may remain unchanged. Review, correction, and exception handling may consume much of the gross saving. Management may remove positions or avoid planned hiring, creating a cost benefit alongside workforce and service consequences.

None of these outcomes is automatically superior. The error is claiming one while managing for another.

If the objective is cost reduction, management must identify what expenditure will be removed or avoided and when. If the objective is capacity, it must define the workload that will absorb the capacity. If the objective is quality, it needs an independent quality measure. If the objective is employee experience, it should measure that experience and watch for work intensification. If the objective is learning, it must name the uncertainty being reduced and the decision that learning will support.

Value is not the conversion of every benefit into dollars. It is the explicit connection between an AI use and an outcome that matters.

■ IN PRACTICE

For every hours-saved claim, record four things: how time was measured, what new work was added, where the released capacity went, and which outcome changed.

1.7 Different levels, different claims

Value can exist at several levels at the same time. At the task level, a person completes an activity faster or better. At the workflow level, a sequence of tasks improves. At the business-unit level, the workflow affects capacity, cost, revenue, quality, customer outcomes, or risk. At the enterprise level, the AI use changes financial performance, strategic capability, resilience, or resource allocation.

A task-level gain is not less real because it has not reached the enterprise. It is simply a different claim.

This resolves an unhelpful argument around AI. One side points to task studies showing substantial gains and concludes that the value question has been settled. The

■ KEY TERM

Level of the claim

Task, workflow, business unit, or enterprise. A gain is real at the level it was measured. It becomes a different claim at the next level.

other points to weak enterprise returns and concludes that the technology is mostly hype. Both can be correct at the level they are measuring.

Management must keep the level of the claim visible. When someone says AI created value, the next question is: for whom, at what level, over what period, and against which cost?

1.8 What the evidence authorizes

A credible value claim needs, at minimum, a named outcome, a baseline, a description of what changed in the work, the material costs and consequences, an accountable owner, a time horizon, a plausible attribution logic, and a decision the evidence is meant to support. The full diagnostic belongs in the toolkit; the management principle belongs here. Early work does not need audited ROI. It does need enough structure that management can distinguish what is known from what is merely hoped.

Do not ask adoption evidence to prove value. Ask what the current evidence legitimately establishes and what next commitment it can authorize.

Once management stops treating adoption as the destination, the next question becomes unavoidable: AI can do many things. Which of those possibilities actually belong in the strategy?

KEY TAKEAWAYS

- Access, usage, adoption, deployment, scale, and value are six different claims. Each requires different evidence.
- Evidence authorizes the decision at its own level. Reversible steps move on limited evidence; hard-to-reverse commitments require stronger proof.
- Hours saved are potential capacity until management decides where the capacity goes and which outcome changes.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *Which of our reported AI metrics has a named owner for the outcome it stands in for?*
- 02 *Where have we asked evidence to authorize a decision beyond its level?*
- 03 *For our headline value claim: for whom, at what level, over what period, and against which cost?*

02

CHAPTER 2 • PART I

Capability Is Not Strategy

THE QUESTION

Which AI possibilities belong in the strategy?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Classify before committing	Effectiveness is not position	Advantage sits in the system around the model	Creating value is not capturing it	Commit at the right moment

DECISION RULE

Capability creates options. Strategy decides which options deserve commitment.

IN THIS CHAPTER

2.1	A catalogue of possibilities is not a direction	27	2.5	Creating value is not the same as capturing it	30
2.2	Operational effectiveness is not strategy	28	2.6	Strategy when capabilities keep changing	31
2.3	The system around the model	29	2.7	What strategy must decide	32
2.4	Model choice becomes strategic when it changes the operating system	30			

MORGAN STANLEY DID not create a strategy by gaining access to GPT-4.

Other financial institutions could buy the same class of model. Many could build a chatbot, summarize documents, or generate meeting notes. The strategic work was in the configuration around the model.

Morgan Stanley Wealth Management connected generative AI to a specific asset: the firm's proprietary research and intellectual capital.¹ It connected that asset to a specific role: helping financial advisors retrieve trusted internal information and serve clients. It built evaluations around real use cases, linked answers to source documents, introduced controls suited to financial services, and later placed its Debrief tool inside a defined client-meeting workflow.²

Company and vendor sources report broad use among advisor teams – evidence of adoption, not audited profit. The strategic point is the placement: the model was not left floating beside the work. It was attached to proprietary knowledge, a trusted distribution channel, an advisor-led service model, operating controls, and an existing customer relationship system.

The model supplied capability. The organization supplied strategic meaning. That distinction is the centre of this chapter.

DECISION

Capability creates options. Strategy decides which options deserve commitment, what they must reinforce, and what the organization will refuse.

2.1 A catalogue of possibilities is not a direction

The most common AI strategy begins as a list. Customer service could use a chatbot. Marketing could generate content. Finance could summarize reports. Human resources could screen documents. Sales could draft emails. Operations could forecast demand. Legal could review contracts. Software teams could generate code.

The list may be accurate. It can still be strategically empty.

A capability map answers: what can the technology do? A strategy answers: which outcomes matter, where will we compete, how will our activities reinforce one another, which assets let us capture value, and what will we refuse or de-prioritize?

AI makes the distinction more important because useful capabilities are increasingly available through shared models, cloud platforms, embedded software, and vendors. A company can gain real efficiency without gaining strategic differentiation. It can also fall behind when an efficiency becomes the new minimum standard.

Management therefore needs to classify AI opportunities rather than celebrate them all as strategic. Some are utilities that improve ordinary work or provide parity. Some are experiments that reduce uncertainty. Some are strategic candidates that reinforce a chosen position and a coherent system of activities. Some should be watched until capability, economics, regulation, or readiness improves. Some should be rejected because they are immaterial, easy to copy, or too costly in risk and complexity.

All five classifications are legitimate. Confusing them produces bad capital allocation.

When a utility is presented as strategy, the organization exaggerates differentiation. When an experiment is treated as strategy, provisional learning hardens into architecture, contracts, staffing, and customer promises before assumptions are tested. When a strategic opportunity is treated as a small productivity feature, management may fail to redesign the business around it.

The first strategic act is classification.

■ KEY TERM

Five classifications

Utility, experiment, strategic candidate, watch, and reject. All five are legitimate. Confusing them misallocates capital.

2.2 Operational effectiveness is not strategy

Michael Porter's distinction between operational effectiveness and strategy remains useful.³ Operational effectiveness means performing similar activities better. Strategy means choosing a distinctive position and building a coherent system of activities that is difficult to match.

Efficiency, quality, and speed are necessary. They are not automatically sufficient because methods that work can be copied, purchased, or embedded into standard software.

Operational AI still matters. A bank that ignores better fraud detection can lose money. A retailer that refuses better forecasting can carry excess inventory. A service company that rejects useful automation can become structurally expensive.

But important and strategic are not synonyms.

BOUNDARY

Important and strategic are not synonyms.

A capability becomes strategically meaningful when it changes the economics or experience of a chosen position and when the surrounding activities reinforce the result. The same model can support opposite strategies. A premium service may use AI to

increase the quality and reach of human advice. A low-cost platform may use it to reduce the need for advice altogether.

The technology does not choose between those positions. Management does.

■ IN PRACTICE

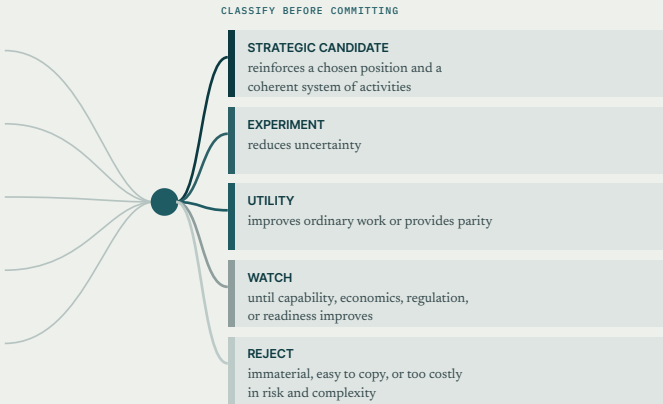
Test for strategy, not importance: does the capability change the economics or experience of the chosen position, and do the surrounding activities reinforce it?

EXHIBIT 2.1

The Strategic AI Test

FIVE STRATEGY QUESTIONS

- 1 Which outcomes matter?
- 2 Where will we compete?
- 3 How will our activities reinforce one another?
- 4 Which assets let us capture value?
- 5 What will we refuse or de-prioritize?



Five questions classify a capability before management makes hard-to-reverse commitments.

SOURCE: AUTHOR.

2.3 The system around the model

The current AI debate increasingly points toward the same mechanism. Productivity improvement alone rarely creates a sustainable advantage when competitors can obtain similar tools. Larger strategic effects appear when AI changes offerings, business models, transaction costs, and the structure of the activity system.

When firms share access to similar models, advantage is more likely to come from privileged data, embedded workflows, scale economics, physical assets, customer relationships, trust, and learning loops. This is not a new law created by AI. It is a current expression of a long-standing strategy principle: an input that everyone can purchase rarely explains durable advantage by itself.

Workflow design matters because production is a sequence, not a collection of isolated tasks. A capability can appear powerful in a demonstration and create little system value

when it is scattered across work full of difficult handoffs. The unit of advantage is often not the model. It is the redesigned system the model makes possible.

The same logic explains why complementary assets matter. A model may be available to everyone. The relevant data may not be. Workflow knowledge may be tacit. Customer trust may have taken decades to build. Distribution may be expensive. Regulatory standing may be difficult to reproduce. A physical network may generate unique information. An installed base may create feedback and learning.

Morgan Stanley's case is strategically interesting for this reason. It placed a commonly available model inside firm-specific research, advisor relationships, and operating controls. It shows how a common capability can be placed inside a more coherent organizational system. Chapter 9 examines the operating configuration in detail.

If the initiative uses the same model, public data, workflow, vendor configuration, and customer interface as competitors, the result may still be valuable. It is simply more likely to be utility than strategy.

2.4 Model choice becomes strategic when it changes the operating system

Model choice belongs inside strategy when it changes the activity system: quality, latency, throughput, review burden, privacy, control, dependency, or cost. The most capable model is not automatically the strategic choice, because management is choosing an operating configuration, not a benchmark winner.

The detailed selection decision comes later, after the workflow and consequence are known. Chapter 9 asks which model, combination of models, or non-AI alternative clears the required performance floor and produces the strongest full economic result. Strategy supplies the purpose and constraints; the model decision supplies the justified configuration.

2.5 Creating value is not the same as capturing it

AI can increase total value in a market without allowing the adopting firm to keep much of it.

A productivity gain may lower the cost of serving customers and competition may pass the benefit into lower prices. A recommendation system may increase conversion while the model vendor, marketplace, data provider, or distribution platform captures much of the economics. A workflow may save employee time while the customer receives the benefit through faster service.

None of those outcomes is automatically bad. They are strategically different.

Management must separate three questions. Where is value created? Who captures it? Who carries the downside?

This is where complementary assets become power arrangements, not items on a checklist. Control of distribution determines who reaches the customer. Control of proprietary data determines who improves fastest. Control of the workflow can create dependence. Contracts can decide who owns the learning produced by use.

A company can create an excellent AI-enabled experience and still weaken its position if a vendor owns the customer interface, retains the data, controls pricing, and can reproduce the capability for competitors. A company can use a common model and strengthen its position when it controls the relationship, integrates the system into distinctive work, and retains the learning generated by the workflow.

The strategic question is not only whether AI creates value. It is whether the organization can capture enough value, maintain acceptable power, and avoid transferring unpriced risk to less powerful stakeholders.

WATCH FOR

Contracts that let a vendor own the customer interface, retain the data, control pricing, and reproduce the capability for competitors.

2.6 Strategy when capabilities keep changing

A reasonable objection remains: how can an organization build strategy around technology that changes every few months?

The answer is not to freeze the strategy around one model.

Model capability is one layer of the system. A durable position can rest on assets and routines that survive model turnover: proprietary data, customer access, distribution, physical operations, regulatory permission, brand, expert knowledge, integrated workflows, and the ability to evaluate and reconfigure quickly.

Speed can create advantage. A firm that learns earlier may accumulate data, customer behaviour, workflow knowledge, or organizational competence before rivals. But speed is not automatically strategic. Moving first into a vendor dependency, broken workflow, or customer promise the organization cannot support can create debt rather than advantage.

Low-consequence utility tools can move quickly under clear rules. Strategic experiments can move quickly when they have a hypothesis, boundary, owner, and stop condition. Difficult-to-reverse commitments – proprietary architecture, workforce restructuring, exclusive contracts, regulated decisions, or customer-rights changes – require stronger evidence.

The strategic asset is not perfect foresight. It is the ability to make coherent commitments at the right moment and revise the system when conditions change.

2.7 What strategy must decide

A strategic AI programme begins with the organization's position and desired outcomes. It distinguishes utility from experimentation and differentiation. It identifies the workflows and assets that must reinforce one another. It states what will be built, bought, protected, or refused. It defines what evidence would justify deeper commitment and what would cause management to stop.

Strong strategy does not mean predicting the future perfectly. It means creating a coherent basis for learning, commitment, and revision.

Capability creates options. Strategy determines which options deserve commitment.

Yet organizations repeatedly reverse that sequence. They buy the tool, announce the programme, and search for the strategy afterward.

That is the tool-first trap.

■ IN PRACTICE

Set the pace by reversibility. Utility tools move under clear rules.

Experiments move with a hypothesis, boundary, owner, and stop condition.

Hard-to-reverse commitments wait for stronger evidence.

KEY TAKEAWAYS

- Classify every AI opportunity – utility, experiment, strategic candidate, watch, or reject – before committing capital.
- Advantage comes from the system around the model: proprietary data, embedded workflows, distribution, trust, and learning loops.
- Creating value and capturing it are separate questions. Control of the interface, the data, and the contract decides who keeps the gain.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 Which of our AI initiatives are utilities presented as strategy?
- 02 Where is value created, who captures it, and who carries the downside?
- 03 What evidence would cause us to stop our most strategic AI commitment?

03

CHAPTER 3 • PART I

The Tool-First Trap

THE QUESTION

Why do organizations buy the tool before they make the operating decision?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Momentum creates facts on the ground	Pilots test the easiest uncertainty	The portfolio tax	The hidden organization around the tool	Procurement is operating design

DECISION RULE

Make the operating decisions before product momentum makes them implicitly.

IN THIS CHAPTER

3.1	The trap is a sequence, not a product category	34	3.5	The hidden organization around the tool	37
3.2	Why organizations begin with the tool	34	3.6	Who gains authority when the tool arrives?	37
3.3	Pilots often test the wrong thing	35	3.7	Shadow AI is demand information	38
3.4	The portfolio tax	36	3.8	Procurement is part of the operating model	38
			3.9	Reverse the sequence	39

A CUSTOMER ASKED AIR Canada about bereavement fares. The airline's website chatbot gave an answer that was wrong. The customer relied on it, later sought reimbursement, and the company argued that the chatbot was a separate legal entity responsible for its own actions. A tribunal rejected that position.⁴ Air Canada remained responsible for information provided on its website.

The striking part of the case is not that a chatbot made a mistake. Models and systems make mistakes. The deeper failure was organizational. A tool had been placed in front of a customer, speaking with the authority of the company, without a sufficiently reliable connection between policy, interface, review, and accountability.

The product existed. The operating decision had not been completed.

BOUNDARY

The tool-first trap is allowing product momentum to make operating decisions that management has not yet made explicitly.

That is the tool-first trap.

3.1 The trap is a sequence, not a product category

Organizations do not fall into the trap merely because they use a vendor product. They fall into it through a sequence.

A capability becomes visible. A senior leader wants momentum. A vendor can demonstrate the feature. A budget exists. The organization acquires or activates the tool. Teams begin experimenting. Only then do questions about outcome, workflow, data, authority, review, measurement, support, and exit arrive.

By that point, the product has created facts on the ground. Employees have built habits. Integrations have started. Contracts have been signed. A public announcement may have been made. The cost of questioning the original decision has risen.

Management has not avoided the hard decisions. It has deferred them until they are more expensive and politically difficult.

3.2 Why organizations begin with the tool

The trap has become subtler as AI is embedded into software organizations already own. A manager may never sign a new AI procurement contract. The capability arrives through a productivity suite, customer platform, analytics product, development environment, or

workflow system. Because the procurement event is invisible, the management event can disappear with it.

Embedded features create a peculiar form of momentum. The vendor has already decided that the capability belongs in the product. The product interface suggests where it belongs in the work. Default permissions suggest who may use it. Usage telemetry suggests what success looks like. Unless management intervenes, product design begins answering questions about placement, authority, evidence, and acceptable consequence that belong to the organization. The absence of a dramatic purchase does not mean the absence of a consequential decision.

The tool is concrete. The operating model is not.

A demonstration can be shown in minutes. A workflow can take weeks to understand. A licence has a price. The cost of review, correction, integration, support, change, and failure is distributed across budgets and time. A vendor has a sales process. The future owner of the business outcome may not yet be in the room.

The tool also creates the appearance of progress. It gives executives something to announce, teams something to test, and employees something to use. Defining the value case, redesigning work, and assigning authority can feel slower because those activities expose disagreement that the demonstration temporarily hides.

The market reinforces the sequence. Vendors sell capabilities they can control and measure. Organizations buy them through procurement systems designed for products and services. The questions that determine business value often sit elsewhere: in operations, finance, risk, customer experience, workforce design, and domain judgment.

Tool-first adoption is therefore not irrational. It is what the existing system makes easiest.

□ WATCH FOR

AI features activated inside software the organization already owns. No purchase event occurs, so no management decision is recorded.

3.3 Pilots often test the wrong thing

A pilot sounds disciplined because it is small. But size alone does not make an experiment useful.

Many AI pilots test whether a model can perform a task under protected conditions. They do not test whether the surrounding workflow can absorb the output, whether users will behave as expected, whether authoritative data is available, whether review capacity is sufficient, whether the economics survive at ordinary volume, or whether the organization can respond when conditions change.

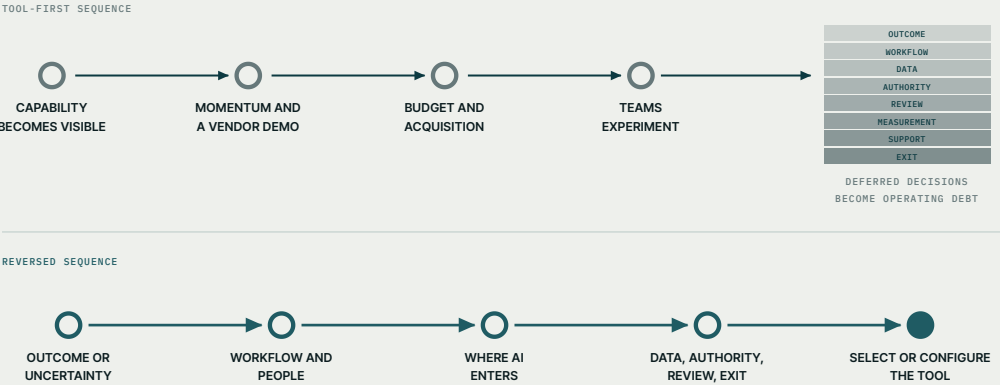
The pilot proves the model can produce something. Management interprets that result as evidence that the AI use can operate.

Those are different claims.

The mistake is not experimentation. It is allowing the easiest uncertainty to test to stand in for the hardest uncertainty management must resolve. Technical feasibility is only one possible unknown. Workflow fit, human behaviour, operating consequence, economics, and scaling burden may matter more to the next commitment.

The most dangerous pilot is not the one that fails visibly. It is the one that succeeds at the wrong question and creates momentum toward a system the organization has never truly designed.

FIGURE 3.1
The Tool-First Trap



Product momentum turns deferred management decisions into operating debt.

SOURCE: AUTHOR.

3.4 The portfolio tax

One tool-first initiative may be survivable. A portfolio of them creates a tax on the entire organization.

The tax becomes visible when every team starts independently. Each group chooses a vendor, negotiates data access, invents review rules, builds a local prompt library, creates its own risk language, and asks the same specialists for bespoke approval. The organization has not merely accumulated use cases. It has accumulated duplicated uncertainty.

That duplication consumes more than money. It consumes security review, legal interpretation, integration capacity, domain expertise, change attention, and executive bandwidth. A portfolio full of individually reasonable pilots can therefore become

collectively irrational because the scarce resource is the organization's ability to make and support the decisions around them.

The organization accumulates duplicate vendors, inconsistent terminology, incompatible controls, and local workarounds. Teams learn different lessons in private. Successful demonstrations wait for integration. Weak pilots remain alive because nobody owns the stop decision. New proposals enter faster than old ones leave.

The visible portfolio is innovation. The hidden portfolio is unresolved obligation.

This is why the value of discipline appears before deployment. A clear outcome, owner, boundary, evidence requirement, and exit condition reduce the cost of learning even when the experiment does not proceed.

■ KEY TERM

Portfolio tax

The duplicated uncertainty created when every team independently chooses vendors, negotiates data access, invents review rules, and requests bespoke approval.

3.5 The hidden organization around the tool

Every useful AI tool creates an organization around itself, whether management designs that organization or not.

Someone chooses the data. Someone decides which outputs are trusted. Someone corrects errors. Someone absorbs exceptions. Someone handles user questions. Someone responds when the vendor changes the model. Someone explains the decision to a customer, regulator, employee, or executive. Someone carries the cost when the system fails.

If those roles are not designed, they do not disappear. They become invisible work.

This is why a licence price can be a poor estimate of cost. The complete business use includes integration, data preparation, policy interpretation, training, review, escalation, incident handling, monitoring, vendor management, change, and the opportunity cost of specialist attention.

The tool is the most visible part of the system. It is rarely the whole system.

3.6 Who gains authority when the tool arrives?

A tool can redistribute authority before anyone formally approves that change.

A ranking system determines what appears first. A summarizer determines which details remain visible. A drafting assistant establishes the first version others edit. A recommender frames the range of plausible action. An agent can convert a suggestion into execution.

The person who selected the tool, configured the prompt, chose the data, or accepted the default may acquire practical influence over decisions owned elsewhere.

That influence is not necessarily improper. It is simply real. Management must decide whether the allocation is justified, visible, and reversible.

The tool-first sequence hides this question because the system is introduced as a productivity feature rather than as a change in who frames, proposes, approves, or acts.

3.7 Shadow AI is demand information

Employees using unapproved AI tools are often treated only as a compliance problem. They are also providing demand information.

Shadow use can reveal repetitive work, poor internal search, missing knowledge infrastructure, slow approval paths, or a capability employees genuinely value. A blanket prohibition may reduce visibility without reducing demand.

The right response is neither indifference nor moral panic. Management should identify what work employees are trying to improve, what data and consequences are involved, which approved path could meet the need, and which uses remain unacceptable.

■ IN PRACTICE

Treat shadow AI as a demand signal. Log the work employees are improving, the data involved, the approved path that could meet the need, and the uses that remain unacceptable.

Shadow AI is evidence that the official operating system has not yet resolved the problem. It is not proof that every unofficial solution should be legitimized.

3.8 Procurement is part of the operating model

Procurement is often treated as the end of selection. For AI, it is part of operating design.

Contracts determine data use, model changes, service levels, audit rights, liability, portability, termination, and who owns the learning generated through use. Vendor dependence can become strategic structure long before the organization calls it strategy.

Procurement therefore cannot end at price, vendor security, and contractual terms. The purchase decision should preserve the organization's ability to define the use boundary, inspect evidence, change permissions, understand material dependencies, retain records, and exit. A contract that makes those decisions difficult can create operating debt before the first user opens the product.

■ IN PRACTICE

Before signing, confirm the contract preserves the right to define the use boundary, inspect evidence, change permissions, see dependencies, retain records, and exit.

A team that cannot explain how it will exit, replace, or reauthorize the capability has not completed the purchase decision.

3.9 Reverse the sequence

The corrective is not to ban tools or require a perfect enterprise design before experimentation. It is to reverse the order of questions.

Name the outcome or uncertainty that matters.

Understand the workflow and the people affected by it.

Decide where AI might enter and what role it would play.

Define data, authority, review, measurement, support, and exit requirements proportionate to the consequence.

Then select or configure the tool that can operate inside those conditions.

This sequence does not eliminate discovery. It makes discovery useful. A reversible experiment can move quickly because management knows what it is trying to learn and what the result will authorize.

The decision rule is simple: make the operating decisions before product momentum makes them implicitly.

■ KEY TERM

The reversed sequence

Outcome or uncertainty → workflow and people → where AI enters → data, authority, review, and exit → only then the tool.

DECISION

Make the operating decisions before product momentum makes them implicitly.

The tool-first trap survives because the visible artifact is easy to confuse with the organizational change. A demo can be purchased. A workflow has to be redesigned. A licence can be distributed. Authority has to be allocated. A model can be benchmarked. The workflow and organization around it have to be made capable of using the result.

This is why transformation cannot be inferred from the presence of AI in the stack. The transformation, when it exists, is in the changed arrangement of work, decisions, incentives, evidence, and capability around the AI use.

A tool enters an organization that already has objectives, incentives, workflows, authority, skills, and feedback. The next question is what the technology will amplify when it arrives.

KEY TAKEAWAYS

- The tool-first trap is a sequence: product momentum answers operating questions before management does.
- A pilot that succeeds at the wrong question creates momentum toward a system nobody designed.
- Procurement is operating design. Contracts decide data use, model change, audit, portability, and exit.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *Which uncertainty did our last pilot actually resolve – and was it the one blocking the next commitment?*
- 02 *What invisible work surrounds our most-used AI tool, and who performs it?*
- 03 *Can each of our AI vendors be exited, replaced, or reauthorized on our terms?*

04

CHAPTER 4 • PART I

AI Scales the Organization You Already Have

THE QUESTION

What will AI amplify when it enters this organization?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
There is no neutral insertion	Six conditions are amplified	Fluency can make weak judgment look complete	Scale hardens conditions into architecture	Commitment must not outrun evidence

DECISION RULE

Examine the organization that will receive the capability before asking what the AI can do.

IN THIS CHAPTER

4.1	The myth of the neutral insertion	42	4.6	Authority and accountability	45	4.11	When scale turns conditions into architecture	49
4.2	Technology and organization are complements	43	4.7	Skill and capacity	46	4.12	The asymmetry of benefit and burden	50
4.3	Objectives and incentives	44	4.8	Feedback and learning	46	4.13	AI can also change the organization	50
4.4	Data and representation	45	4.9	AI can make weak judgment look complete	46	4.14	The management consequence	51
4.5	Workflow and handoffs	45	4.10	When apparent competence receives too much credit	47			

A WIDELY USED HEALTHCARE algorithm was designed to identify patients who would benefit from additional care. The system did not use illness directly as its principal proxy. It used healthcare cost.

That choice appeared reasonable. Cost was measurable, available, and correlated with need. But Black patients with the same level of illness had historically generated lower healthcare spending because of unequal access and treatment. The algorithm therefore interpreted lower cost as lower need.

Correcting the proxy would have increased the share of Black patients receiving additional help from 17.7 per cent to 46.5 per cent.¹

The algorithm did not invent the inequality. It learned from an operating system in which spending, access, and need were already misaligned. Then it converted that misalignment into a repeatable decision rule.

This is the deeper meaning of the claim that AI scales the organization you already have.

AI enters objectives management has chosen, data the organization has produced, workflows people already follow, incentives that reward some outcomes over others, and authority structures that determine whose judgment matters. It can improve those systems. It can also make their weaknesses faster, cheaper, more consistent, and harder to see.

■ EVIDENCE

The proxy – healthcare cost – encoded unequal access. Correcting it would have raised the share of Black patients receiving additional help from 17.7 to 46.5 per cent.

BOUNDARY

AI does not enter a neutral organization. It amplifies the objectives, representations, workflows, authority, skills, and feedback systems into which it is placed.

4.1 The myth of the neutral insertion

Organizations often speak as though AI is inserted into a stable process and contributes a bounded capability. In practice, the insertion changes both the work and the organization around it.

The AI may alter what information becomes visible, which cases receive attention, how people allocate time, what employees learn, which errors are recorded, and who gains practical authority. The organization also shapes the AI through data selection, system configuration, prompts, policies, incentives, and user behaviour.

The result is reciprocal. Technology changes the organization. The organization changes what the technology becomes in use.

That is why the same model can produce different consequences across firms. One organization surrounds it with clear objectives, authoritative data, capable users, disciplined review, and rapid learning. Another surrounds it with conflicting incentives, fragmented ownership, weak data, and no meaningful correction path. The model may be identical. The way it is used inside the organization is not.

4.2 Technology and organization are complements

Research on information technology has repeatedly shown that technology and organizational design are complements.² Investments in software and computing create larger benefits when firms also change skills, decision rights, processes, incentives, and coordination.

AI intensifies that relationship because it acts closer to judgment and work. It does not merely store or transmit information. It can generate, classify, rank, recommend, and execute. The surrounding organization therefore affects not only how much value appears, but what kind of consequence the system produces.

The complementarity point is easy to understate. A model is not added to an organization the way a faster processor is added to a machine. Its effect depends on what other elements change with it. Workflow, data, incentives, roles, training, targets, review capacity, and decision rights can increase or destroy the value of the same technical capability.

This is also why complementary investment is not secondary implementation work. It is part of the economic asset being built. The software may be purchased in a quarter; the capability to use it well is accumulated through redesigned work, new routines, better evidence, and people who know when the system should not be trusted.

The lesson is not a generic call to focus on people and process. It is more demanding: management must identify the specific organizational conditions the capability will rely on and the weaknesses it may scale.

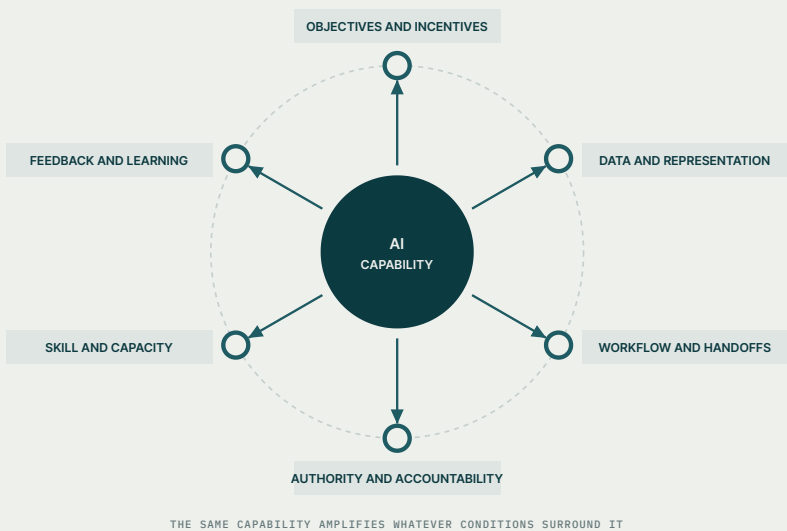
4.3 Objectives and incentives

AI optimizes what the organization makes legible. If management defines success through the wrong proxy, the system can become excellent at producing the wrong result.

■ IN PRACTICE

Name the proxy the system will optimize. Then describe what it would look like to hit the proxy while missing the outcome – and monitor for exactly that.

FIGURE 4.1
What AI Amplifies



A causal lens showing why the same capability produces different results in different organizations.
SOURCE: AUTHOR.

A customer-service system rewarded only for handling time may shorten conversations while increasing repeat contact or customer frustration. A sales assistant rewarded only for conversion may make promises operations cannot fulfil. A hiring model trained on historical success may reproduce a narrow definition of who appears promising.

The model does not decide what the objective should be. The organization does, whether explicitly or through the metrics it supplies.

This is why value definition is not an accounting step at the end. It is part of system design at the beginning.

4.4 Data and representation

Data is not a neutral record of reality. It is a record of what the organization collected, how categories were defined, which interactions occurred, and which outcomes were measured.

Historical data can encode exclusion, inconsistent service, missing populations, workarounds, and policies that no longer reflect the organization's intent. When AI learns from that record, it can convert contingent history into apparent objectivity.

The question is not only whether the data is clean. It is whether the representation is appropriate for the decision being made.

4.5 Workflow and handoffs

A task improvement can make the surrounding workflow worse. Faster generation can create more material for someone else to review. Better triage can flood a specialist queue. Automation at one step can increase exceptions at the next. A local team can improve its metric while transferring cost and delay downstream.

AI therefore scales the structure of work, not only the performance of a task. If the workflow is fragmented, the technology may move the bottleneck rather than remove it.

The relevant unit is the end-to-end outcome, including the handoffs and queues the AI use changes.

4.6 Authority and accountability

AI can redistribute authority without changing an organization chart. The system that selects what appears first can frame the decision. The recommendation that arrives before human analysis can anchor judgment. The default that is easiest to accept can become practical policy.

If nobody is explicitly authorized to challenge the frame, correct the proxy, suspend the system, or remedy the consequence, the organization may retain formal accountability while losing practical control.

Accountability must therefore follow the consequence, not merely ownership of the software.

DECISION

Accountability must follow the consequence, not merely ownership of the software.

4.7 Skill and capacity

AI can spread expertise. In some settings, less experienced workers gain more because the system makes high-quality patterns available at the moment of work.³

But fluency can hide an inadequate frame, missing context, or the wrong standard. The people least equipped to recognize the omission may be the most likely to accept fluency as evidence of expertise.

Reviewer competence is therefore more than knowing whether a sentence is correct. It includes knowing what good looks like well enough to detect when the system has answered the wrong question.

Capacity matters too. A qualified reviewer with no time, no evidence, or no authority is not an effective control. The organization can claim that a human remains in the loop while designing conditions that make meaningful review impossible.

□ WATCH FOR

Claims that a human remains in the loop when the reviewer lacks the time, evidence, or authority to change the result.

4.8 Feedback and learning

AI systems operate in changing environments. Data shifts. Policies change. Users adapt. Vendors update models. Customers learn how to interact with the system. A result that was acceptable at launch can become unreliable or uneconomic later.

An organization with strong feedback can detect the change, investigate the cause, and revise the operating decision. A weak organization records usage while losing sight of outcomes. It may scale the system precisely as the original conditions stop holding.

The quality of the learning loop determines whether AI improves the organization or simply accelerates it.

4.9 AI can make weak judgment look complete

The obvious AI failures are often easier to manage because they announce themselves. An invented fact, a broken integration, or an impossible calculation creates friction. Someone

notices that something is wrong. A fluent answer built inside the wrong frame is more dangerous precisely because it may create no such interruption.

The answer can be factually plausible, logically tidy, balanced in tone, and institutionally acceptable while leaving the governing assumption untouched. It can explain the visible actor without examining the incentive system around that actor. It can optimize the stated proxy without asking whether the proxy is what the organization should value. It can summarize the approved evidence without noticing that the approved evidence excludes the stakeholder carrying the downside.

That is not a problem that better prompting alone can solve. Someone has to know what a good analysis should contain. Someone has to recognize when the unit of analysis is too narrow, when an inherited category is doing hidden work, when a missing comparison changes the conclusion, or when apparent neutrality simply reproduces the organization's default frame.

This is one reason I do not treat human judgment as a final approval step at the edge of the system. Judgment sits upstream in problem definition and downstream in interpretation, remedy, and change. The organization must remain capable of thinking above the output it receives.

AI does not automatically raise an organization to the level of the model's capabilities. Without a stronger system of judgment around it, AI can scale the level at which the organization already knows how to think – and make that level look more complete than it is.

4.10 **When apparent competence receives too much credit**

A system can produce an answer that looks more knowledgeable than the evidence underneath it deserves. Fluency is part of that effect, but not the whole of it. Citations, confident structure, retrieval from authoritative documents, polished interface design, memory, voice, tool use, and human correction can all make the system appear to understand a problem more completely than it actually does. The result may be genuinely useful. The management error begins when usefulness is converted into an unwarranted attribution of understanding or authority.

I call this Engineered Epistemic Performance: the socio-technical production of knowledge-shaped language, behaviour, and interface signals that can cause users or organizations to attribute understanding, completeness, or authority beyond what the available evidence warrants. It sits upstream of the AI Value Gap because it helps explain why capability can receive too much credit before management has established whether that capability produces worthwhile value.

The mechanism is familiar in other forms of automation. Human-factors research has long shown that trust, interface cues, workload, reliability, and the way automation is presented can change reliance and monitoring.⁴ Research on language models has separately warned against confusing linguistic performance with demonstrated understanding.⁵ More recent chatbot research shows that perceived intelligence and human-like qualities can influence competence trust and willingness to adopt.⁶ Together these literatures establish the management point: the appearance of knowing changes how people treat a system.

The sequence matters because organizational authority can accumulate without a formal delegation. Knowledge-shaped output and interface cues create perceived understanding or completeness. Perceived competence can reduce scrutiny or make the generated recommendation the anchor for review. Repeated acceptance gives the system practical authority inside the workflow.

Practical authority then becomes investment, staffing, integration, policy, customer consequence, or some other organizational commitment.

This is different from hallucination. A false statement is one possible failure mode, and often an easier one to detect. Engineered Epistemic Performance concerns the system-level appearance of knowing even when individual statements are correct. A recommendation can cite the right policy and still use the wrong frame. A summary can accurately reproduce the approved evidence while omitting the evidence management never asked the system to retrieve. A technically correct answer can therefore remain managerially incomplete.

The management response is not to assume deception or to distrust fluent systems by default. It is to separate the sources of perceived capability and ask what evidence warrants the authority being granted. Which part of the performance comes from the underlying model? Which part comes from retrieval, curation, prompting, tools, interface cues, selective exposure to easy cases, or human correction? What independent standard would show that the apparent competence is sufficient for this decision and consequence? That is the point at which judgment becomes an operating requirement rather than a ceremonial human sign-off.

■ KEY TERM

Engineered Epistemic Performance

Knowledge-shaped language, behaviour, and interface signals that lead people to attribute more understanding, completeness, or authority to a system than the evidence warrants.

4.11 When scale turns conditions into architecture

At small scale, organizations can hide weak conditions behind exceptional people. A highly motivated pilot team remembers the edge cases, knows whom to call, manually checks the output, and repairs failures before customers notice. Scale removes that protection. Volume turns personal memory into queues, informal judgment into inconsistent policy, and heroic support into an operating cost.

The question at scale is therefore not whether the model still works. It is whether the organization has converted tacit rescue work into explicit capability: reliable data, routable exceptions, trained reviewers, delegated authority, current records, and feedback that changes the system.

Scale therefore hardens conditions into architecture. Weaknesses that were manageable in a protected pilot become recurring costs, risks, and obligations.

This is why scale is not simply more users. It is the point at which the organization must prove that the surrounding system can carry the capability without exceptional protection.

■ IN PRACTICE

Before scaling, list the rescue work the pilot team performs by hand. Each item must become an explicit capability – or it is a reason not to scale yet.

CONTEXT · EVIDENCE

Meta's Project OT: commitment ahead of evidence

Meta's Project OT provides a contemporary example of that problem at organizational scale. Reuters reported that the company considered an "AI native" operating model in which smaller, talent-dense human teams would oversee AI-supported work.⁷ Scenario planning examined reductions of up to 60 per cent in some teams, alongside redeployment, role closures, and layoffs. Meta carried out a roughly 10 per cent workforce reduction in May 2026, while a planned second company-wide wave was cancelled before it occurred.

The operating evidence reported alongside the restructuring is more revealing than the headline about jobs. Internal measures cited by Reuters showed code changes to internal platforms and infrastructure rising far faster than changes reaching users – 220 per cent year over year against 36 per cent. The same reporting described increases in major technical and security incidents and in firefighting time, while employee favourability fell. Meta later paused a mouse-tracking programme, allowed some transfers back, sought to restore stability, and Mark Zuckerberg acknowledged that agent technology had not accelerated as quickly as he had expected.

Whatever caused each operational, security, workforce, and morale outcome, the management lesson is specific: a hard-to-reverse commitment can begin carrying an

assumption about future capability before the surrounding workflow has demonstrated that capability under ordinary load.

Several failures can then hide inside one success story. More generated or changed code can be mistaken for more user value. Review, remediation, security response, coordination, and firefighting can become a second workflow that gross production measures do not capture. Role ambiguity and weak organizational permission can damage the workflow and organization around the technology. And a capability forecast can become embedded in workforce structure before the forecast has earned that weight.

The corrective actions matter as well. Cancelling a further wave, narrowing tracking, permitting transfers, and acknowledging slower-than-expected capability are evidence that management changed parts of the commitment rather than treating the original plan as irreversible. The value-management principle is not “never restructure around AI.” It is that hard-to-reverse commitments should not outrun the evidence that authorizes them.

4.12 The asymmetry of benefit and burden

AI benefits and burdens rarely land in the same place. One team may receive faster output while another receives more review work. The company may save cost while customers carry error. A vendor may capture revenue while the organization retains accountability. Senior leadership may receive credit for speed while frontline employees absorb correction and escalation.

A local dashboard can therefore improve while enterprise value declines. Management has to follow the consequence across the full business use, not stop where the most convenient metric lives.

A complete value case therefore asks whose outcome improves, who carries downside, and who has the authority to interrupt the system.

Distribution is not a side issue to be handed to governance after the business case is approved. It is part of the business case.

4.13 AI can also change the organization

The relationship is reciprocal over time. Once people begin relying on an AI system, they change what they practice, what they notice, which knowledge remains local, and which decisions are escalated. The AI use can create new expertise while allowing other expertise to decay.

Management should therefore ask not only what the organization does to the AI system, but what repeated use of the system is doing to the organization. A capability that improves today’s throughput while weakening tomorrow’s judgment, resilience, or learning may be creating a liability that ordinary productivity metrics cannot see.

The argument is not that organizations are trapped forever by their current conditions. AI can reveal bottlenecks, expose inconsistent policy, spread useful expertise, and create pressure to clarify authority. It can become a catalyst for better work.

But that transformation is a management achievement, not an automatic property of the model.

The organization must decide what should change around the capability: the objective, workflow, role, skill, data, control, incentive, or feedback loop. Without that work, AI is more likely to amplify the current system than to repair it.

4.14 The management consequence

Before asking what the AI can do, examine the organization that will receive it. What objective is the system likely to optimize? What history does the data represent? Where will the bottleneck move? Who gains practical authority? Who knows what good looks like? Who carries the burden? What evidence will reveal whether the conditions have changed?

These questions explain why the management system is necessary.

The first four chapters have removed four illusions: adoption is not value, capability is not strategy, a tool is not an operating system, and AI does not enter a neutral organization.

The management layer is now visible. The next task is to define it – and to identify the minimum conditions under which capability can become worthwhile value.

KEY TAKEAWAYS

- AI amplifies the objectives, data, workflows, authority, skills, and feedback it enters.
- Fluent output can make a weak frame look complete. Judgment belongs upstream and downstream of the output, not only at sign-off.
- Scale converts rescue work into architecture. Hard-to-reverse commitments must not outrun the evidence.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *What proxy will this system optimize, and how could it hit the proxy while missing the outcome?*
- 02 *Who carries the burden of this AI use, and who has the authority to interrupt it?*
- 03 *What is repeated use of this system doing to our judgment and expertise?*



PART II

The Management System for AI Value

Build the management system that chooses a worthwhile result, connects the conditions around one business use, and makes evidence govern what happens next.

05	AI Value Management and the Five Conditions	54
06	Define the Value Case Before the Use Case	70
07	From Output to Outcome	83

05

CHAPTER 5 • PART II

AI Value Management and the Five Conditions

THE QUESTION

What minimum system turns one AI use into worthwhile value?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Manage one intervention in real work	Six recurring decisions	Five Conditions of AI Value	Three connected evidence loops	The weakest condition governs

DECISION RULE

The burden should match the decision being authorized.

THE FIVE CONDITIONS • WHERE THIS CHAPTER SITS

Value clarity	Workflow fit	Operating rules	Ownership and accountability	Evidence and learning
---------------	--------------	-----------------	------------------------------	-----------------------

IN THIS CHAPTER

5.1	The failure appears between functions	55	5.6	Why five conditions	58	5.13	Condition 5: Evidence and learning	65
5.2	The managed unit is one business use inside real work	56	5.7	The system at a glance	59	5.14	Evidence must be able to change the intervention	66
5.3	A discipline of connected decisions	56	5.8	One intervention, five conditions	60	5.15	The weakest condition governs the decision	66
5.4	The discipline in motion	57	5.9	Condition 1: Value clarity	61	5.16	Proportionality without framework theatre	67
5.5	What the discipline connects	57	5.10	Condition 2: Workflow fit	62	5.17	The management standard	68
			5.11	Condition 3: Operating rules	63			
			5.12	Condition 4: Ownership and accountability	64			

THE PROBLEM WAS not a shortage of functions or expertise.

Organizations already had strategies, technology teams, risk policies, finance processes, project offices, data functions, change programmes, and steering committees. Each could perform legitimate work. The persistent failure appeared between them.

I kept returning to the same picture: the strategy team could explain why AI mattered; technology could explain whether the system ran; risk could explain what might go wrong; finance could challenge the forecast; operations could describe the process; a vendor could report adoption and model performance.

Yet the organization could still be unable to answer one practical question.

Is this specific business use of AI creating worthwhile value, and what should management do next?

That question crosses the life of the use. It begins before a tool is selected and remains open after launch. It connects the outcome, workflow, permissions, ownership, economics, evidence, and ability to change course.

This chapter gives that work a name.

5.1 The failure appears between functions

Management disciplines appear when an important result can no longer be produced reliably through scattered judgment.

Quality management does not mean that one department manufactures every product. It connects objectives, process, responsibility, evidence, and improvement so that quality can be produced repeatedly. Innovation management does not invent every idea. It creates a system for selecting, supporting, testing, and learning from innovation. Benefits realization keeps expected benefits visible beyond project delivery.

AI now creates a similar coordination problem.

A single business use can begin as a technical capability, enter a workflow, change an employee role, affect a customer, consume variable infrastructure, depend on a vendor, create new review work, and generate evidence across several systems. The result can change after launch because the model, data, process, price, policy, user behaviour, or market changes.

The use is simultaneously technical, operational, financial, organizational, and consequential. Managing only one dimension leaves the others to chance.

Existing disciplines remain necessary. Their decisions need a shared object, a shared record, and a way to authorize the next move.

5.2 The managed unit is one business use inside real work

Organizations often manage AI at the wrong level.

At the highest level, they manage an enterprise AI strategy. The strategy identifies ambitions, domains, platforms, investment priorities, and principles. That is necessary, but too broad to determine what should happen when one output is wrong, one queue grows, or one customer journey stops creating value.

At the lowest level, teams manage a model, prompt, feature, agent, or integration. That is also necessary, but technical performance alone cannot determine whether the organization should continue paying for, relying on, or expanding the system.

AI Value Management works at the level between them: a defined business use of AI inside a defined system of work.

This book calls that complete managed unit an AI-enabled intervention. The term is useful when the discussion must include more than the technology: the workflow, people, data, decisions, permissions, costs, controls, affected stakeholders, evidence, and current authorization.

■ KEY TERM

AI-enabled intervention

One defined business use of AI inside a defined system of work: its workflow, people, data, decisions, permissions, costs, controls, stakeholders, evidence, and current authorization.

A meeting summarizer used privately by one employee is one intervention. An assistant retrieving approved research for financial advisors is another. An agent changing a customer account is another. A model recommending credit, maintenance, medical attention, or hiring review is another.

The technology category may be similar. The management object is different because the outcome, authority, reversibility, and consequence differ.

This level prevents two errors. Managing AI only as an enterprise programme can make the portfolio look coherent while individual workflows remain unowned. Managing AI only as a technical asset can make the model look healthy while the business result deteriorates.

The intervention keeps strategy above it and operating evidence below it connected.

5.3 A discipline of connected decisions

A phrase does not become a discipline because it sounds important.

A credible discipline needs a defined object, recurring decisions, accountable roles, evidence, and a feedback loop.

The object is the AI-enabled intervention. The recurring decisions concern what to pursue, how to design it, what to permit, whether the organization can operate it, what

evidence counts, and whether to expand, change, restrict, renegotiate, or stop.

The roles remain distributed. A business owner carries the result while technical, financial, risk, workforce, and domain specialists own decisions within their competence. The evidence must distinguish availability, technical performance, use, workflow outcome, financial consequence, and stakeholder effect rather than combining them into one success story.

The feedback loop matters because value is not secured in an approval document. The intervention must be reassessed as the model, workflow, cost, user behaviour, and external conditions change.

Without those properties, AI Value Management is a slogan attached to a transformation office. With them, it becomes an operating discipline.

AI Value Management is the discipline of selecting, designing, authorizing, operating, proving, and adapting specific business uses of AI so that capability becomes worthwhile, measurable, and sustainable value.

I am not proposing that every organization create another office. I am proposing that management stop allowing one business result to dissolve across separate strategies, projects, policies, dashboards, and vendors. The work already exists. The discipline connects it around one outcome and one decision chain.

5.4 The discipline in motion

The discipline moves through six recurring decisions. Select the outcome worth changing and determine whether AI is a credible option. Design the intervention inside the real workflow. Authorize what the system may do and who may challenge or stop it. Operate it under ordinary conditions rather than protected pilot conditions. Prove what changed after full cost and consequence are counted. Adapt the commitment as evidence, technology, cost, demand, and context change.

These are not ceremonial gates. Low-consequence and reversible uses can move quickly. Consequential commitments require stronger evidence. The purpose is to keep learning connected to authority.

5.5 What the discipline connects

AI strategy decides where the organization intends to compete and build capability. Operations owns the work. Technology and data teams make the system perform. Finance challenges cost and realization. Risk, legal, privacy, security, and governance define mandatory boundaries. Product, workforce, and domain specialists shape how the intervention affects people and decisions.

AI Value Management does not replace those disciplines. It gives them a shared management object and forces their evidence into one coherent decision. Its distinct job is to keep the business reason, workflow, authority, economics, operating rules, evidence, and ability to change course attached to the same version of the same use.

5.6 Why five conditions

I arrived at the Five Conditions by repeatedly asking what had to be present before a technically impressive use could become a defensible management commitment. The answer was never one more specialist function. Strategy without workflow could not place the capability. Workflow without operating rules could not contain it. Rules without ownership could not make anyone answerable. Ownership without evidence could not tell management whether to continue. Evidence without a worthwhile outcome could prove the wrong thing very efficiently.

The point of reducing the system to five conditions is not elegance for its own sake. It is to give management a minimum set of connected questions that can survive changes in vendors, models, sectors, and organizational structure. The labels matter less than the discipline of refusing to let one missing condition disappear inside an average.

Conditions, not departments or stages

The word conditions is deliberate.

A condition is a state that must be present for an outcome to become likely. It is not a department, document, or meeting. It cannot be satisfied merely because someone completed a template.

Value clarity is not a finance workstream. Workflow fit is not an operations workstream. Operating rules are not a risk workstream. Ownership is not an org-chart box. Evidence is not a dashboard. Each condition crosses functions.

The Five Conditions are not a five-step implementation sequence or five maturity levels. An organization may map the workflow, discover that the value claim is weak, and return to the outcome. A risk review may expose a missing owner. Runtime evidence may force a rule change. The conditions interact and must be revisited when the intervention changes.

They are closer to a conjunctive test than an additive score. A customer-facing agent with excellent workflow fit and no accountable owner is not four-fifths ready. A highly governed system with no worthwhile outcome is not partly valuable. It is not valuable at all.

□ WATCH FOR

Composite scores. A strong average can hide the one missing condition that makes the use unmanageable.

The organization does not need perfection. Each condition must clear the threshold required by the consequence, reversibility, scale, and uncertainty of the decision being authorized.

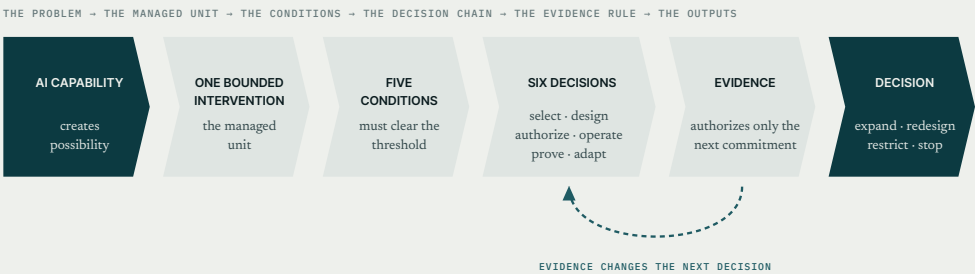
The framework synthesizes durable ideas from benefits realization, workflow and sociotechnical design, management systems, AI risk and impact practice, measurement, and continuous monitoring.¹ The contribution is a minimum management model for one intervention: the smallest connected set of conditions management needs in order to decide whether capability should become operating value.

5.7 The system at a glance

The book can be entered through one simple chain. AI capability creates possibility. Management defines one bounded intervention, establishes the conditions that make value possible, makes a sequence of explicit decisions, and lets evidence authorize only the next justified commitment. The detailed instruments sit in the companion toolkit; the narrative keeps the decision logic visible.

FIGURE 5.0

AI Value Management at a Glance



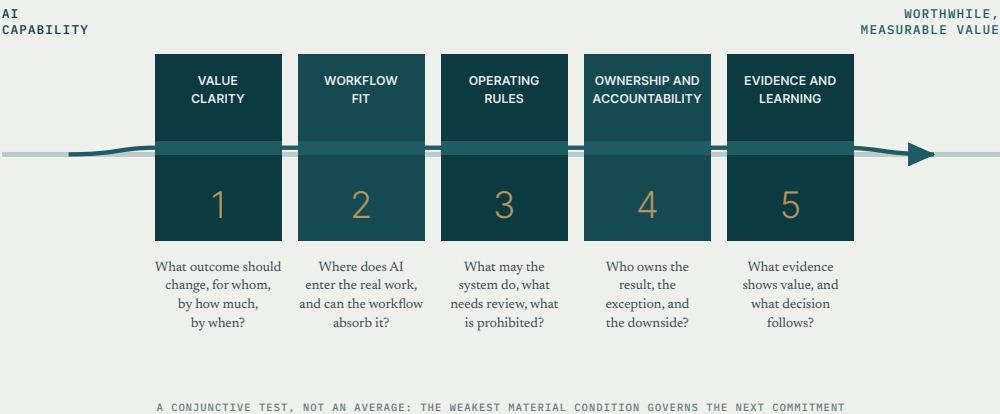
The problem, managed unit, Five Conditions, decision chain, evidence rule, and decision outputs in one view.

SOURCE: AUTHOR.

The figure is a conversion system, not a funnel. AI capability enters from the left. Measurable value appears only when the conditions around the use are strong enough to convert capability into an outcome the organization can operate, govern, measure, and improve. The sequence helps explanation; the conditions themselves remain interdependent.

5.8 One intervention, five conditions

FIGURE 5.1
The Five Conditions of AI Value



One connected minimum system for converting capability into worthwhile value.

SOURCE: AUTHOR.

The Five Conditions function as a contract between disciplines that usually see different parts of the same system.

Consider the illustrative billing-support assistant used throughout the next two chapters.²

The proposed use is not simply to deploy generative AI for billing questions. It is to retrieve approved billing policy and draft an answer for an agent to review. The intended outcome is to reduce repeat billing contacts from 31 per cent by at least five percentage points within six months without worsening resolution quality, complaints, escalation, adjustment errors, or specialist backlog.

Value clarity defines the outcome, beneficiary, baseline, time horizon, material threshold, and unacceptable side effects.

Workflow fit locates the assistant after the agent identifies the billing issue but before the response is sent. It makes the policy source, agent review, difficult-case routing, specialist queue, and downstream adjustment process visible.

Operating rules allow the system to retrieve approved policy and draft a response. It may not change an account, issue an adjustment, invent an exception, or send a customer message without human approval. Low-confidence or excluded cases route to a specialist.

Ownership and accountability place the business result with the customer-service executive. The workflow manager owns process change. Technology owns system reliability and observability. Policy owners maintain authoritative content. Risk and privacy define mandatory boundaries. Finance validates the economics.

Evidence and learning connect source grounding, response quality, agent edits, handling time, repeat contacts, escalations, complaints, specialist workload, full cost, and customer effort. Thresholds determine whether the organization expands, redesigns, restricts, or stops.

The example is illustrative, not a reported company case. Its purpose is to show how one use becomes one management object.

The conditions should therefore appear in one versioned record for the AI use rather than five unrelated documents. Otherwise, finance can validate an early business case, risk can approve a bounded pilot, technology can deploy a changed model, and operations can scale a modified workflow while the dashboard reports them as though they still describe the same system.

DECISION

A material change in one condition must propagate through the others.

5.9 Condition 1: Value clarity

What worthwhile outcome should change, for whom, by how much, and by when?

This is stricter than naming a use case.

“Use generative AI for customer service” describes a location for technology. “Reduce repeat contacts for billing questions without lowering resolution quality or transferring more work to escalations” describes an outcome that management can examine.

Management must name the outcome. It must identify the beneficiary. It must establish the current baseline. It must choose a time horizon. It must define a materiality or decision threshold. And it must state which costs, risks, or disbenefits could make an apparent gain unacceptable.

The beneficiary matters because value is not automatically shared.

An automation can reduce cost for the company while making a remedy harder for customers. It can increase output for one team while moving verification work to another. It can improve speed for managers while increasing surveillance or uncertainty for employees. The intervention may still be worthwhile, but the distribution of benefit and burden belongs inside the value claim.

The baseline matters because improvement cannot be demonstrated against a memory of the old process.

The threshold matters because a statistically detectable change may not justify the integration, operating burden, risk, or opportunity cost. A pilot can improve a metric and still not deserve scale.

Value clarity also protects experiments from an unfair demand for immediate profit. Learning, option value, capability building, quality, resilience, and customer experience can be legitimate outcomes. They need the same discipline: a named uncertainty, milestone, owner, and decision that the learning will support.

Without value clarity, every positive number becomes a candidate success metric.

Usage rises. Response time falls. Employees report saved hours. Costs increase. Quality moves in both directions. Management has data but no rule for deciding whether the intervention is working.

Minimum proof for this condition is not an ROI percentage. It is a value hypothesis precise enough to be tested.

■ IN PRACTICE

Write the value hypothesis in one sentence: outcome, beneficiary, baseline, threshold, time horizon, and the side effects that would make the gain unacceptable.

5.10 Condition 2: Workflow fit

Where precisely does AI enter the real work, and can the surrounding workflow absorb what it does?

A capable model can be badly placed.

It may produce an answer after the moment when the decision is made. It may automate the easiest task while leaving the bottleneck untouched. It may increase volume into a downstream queue. It may require information the workflow does not reliably capture. It may create more verification than the task it removes.

Workflow fit therefore examines the full chain.

What happens before the AI acts? Which inputs are authoritative? Which tasks are adjacent? Who receives the output? What decision follows? Which exceptions break the normal path? What work is created for review, repair, explanation, support, and remedy? What changes when volume increases?

MIT's task-chain research is useful here because it moves analysis beyond isolated task performance.³ AI-friendly tasks can create greater value when they sit beside one another and can be reorganized into coherent chains. Repeated human-machine handoffs, difficult boundary tasks, and downstream constraints can erase the apparent gain.

The correct unit of design is therefore not the prompt or model response. It is the workflow that produces and consumes the response.

Workflow fit is not the same as employee convenience.

A tool can be easy to use and poorly integrated. Employees may copy information between systems, reconstruct missing context, or maintain unofficial checks. The intervention appears adopted because people have learned to compensate for it.

That compensation is evidence about the workflow.

The strongest design work happens close to the people who understand the actual process, including the workarounds and exceptions absent from formal documentation. Their role is not to approve technology after it is selected. It is to reveal what the system must survive.

Minimum proof includes a mapped current and future workflow, named handoffs and exceptions, an estimate of new human work, and evidence that the intervention reaches the actual constraint rather than an attractive nearby task.

5.11 Condition 3: Operating rules

What may the system do, what requires review, what is restricted, and what is prohibited?

Rules convert general principles into operating behaviour.

A responsible-AI policy may state that humans remain accountable. The workflow still needs to specify which outputs require review, what the reviewer checks, what evidence is available, how much time the reviewer has, and what happens after disagreement.

A data policy may prohibit sensitive information from entering unapproved systems. The workflow still needs to define which fields, documents, tools, and user actions are allowed.

An autonomy principle may require human control. The intervention still needs to specify which actions the agent can propose, prepare, execute, reverse, or never perform without permission.

Operating rules should cover at least four boundaries.

The first is information: what data may enter, leave, be retained, or be used for learning.

The second is action: what the system may generate, recommend, decide, communicate, or execute.

The third is review: which conditions trigger validation, approval, escalation, override, or suspension.

The fourth is consequence: what remedy, incident response, recovery, and shutdown process exists when the system is wrong or the environment changes.

Good rules do not make every use slow.

They create differentiated lanes. Low-consequence, reversible work can move with light controls. Consequential actions receive stronger review, evidence, and authorization.

Clear boundaries reduce repeated negotiation and make safe speed possible.

Bad rules fail in two directions.

They can be so vague that every employee interprets them differently. Or they can be so broad and restrictive that people route around them through shadow use.

The rule set must fit the actual capability and consequence. “Human in the loop” is not an operating rule. “A licensed advisor must review the source-linked recommendation before it reaches a client and may edit, reject, or escalate it” is closer.

■ IN PRACTICE

Test each rule at the point of work: can a frontline employee tell, in the moment, what is allowed, what needs review, and what is prohibited?

Minimum proof includes allowed, restricted, and prohibited uses; data and action boundaries; review and escalation thresholds; override and shutdown rights; and a path for remedy.

5.12 Condition 4: Ownership and accountability

Who owns the business result, the operating decision, the exception, and the downside?

AI initiatives often have many participants and no accountable owner.

Technology owns the platform. A product team owns the feature. Risk owns the policy. Operations owns the process. Finance validates the business case. A vendor changes the model. Frontline employees handle errors. Customers experience the consequence.

The organization can distribute work. It cannot distribute accountability until it disappears.

The intervention needs one accountable business owner with authority over the outcome and enough access to coordinate the functions that shape it. That person does not personally perform every task. The role exists so that trade-offs do not remain trapped between departments.

Ownership also needs a decision-rights map.

Who may propose the intervention? Who approves deployment? Who operates it? Who sets thresholds? Who reviews exceptions? Who can override an output? Who can suspend the system? Who decides whether evidence permits scale or requires termination? Who communicates with affected people?

Accountability is incomplete when the person carrying responsibility lacks authority or evidence.

A frontline employee cannot be made accountable for an automated recommendation if the system hides the basis, the employee has seconds to respond, and overriding it creates personal risk. A business unit cannot own value if Finance controls the baseline, Technology controls the data, and a vendor can change the system without a coordinated decision.

The power arrangement must match the accountability claim.

This condition also identifies who carries the downside. The executive receiving credit for cost reduction may not see the customer appeals, employee review burden, or technical debt that makes the result possible. An accountable owner must have a view across those boundaries.

Minimum proof includes a named outcome owner, clearly assigned operating and control roles, explicit override and stop authority, and a route through which affected people can surface evidence that the system is failing.

5.13 Condition 5: Evidence and learning

What evidence will show whether value is being created, whether the system remains acceptable, and what decision follows?

Evidence begins before deployment.

Management needs a baseline, a testable hypothesis, and criteria for the next decision. A pilot that produces interesting output but cannot authorize scale, redesign, restriction, or termination has generated activity rather than management evidence.

Evidence also operates at several levels.

Technical measures show whether the model and infrastructure perform. Usage measures show whether people engage. Workflow measures show what happens to throughput, quality, queues, exceptions, and human effort. Outcome measures show whether the intended customer, financial, workforce, risk, or strategic result changed. Cost measures show the full resources consumed. Impact evidence shows who benefited, who was burdened, and whether new harms appeared.

The levels must be connected.

A model can improve while the workflow deteriorates. Usage can rise while value falls. An operational metric can improve while full cost increases. A financial benefit can be real while an unacceptable consequence is shifted onto a less powerful stakeholder.

The fifth condition therefore includes monitoring, not merely measurement.

The intervention needs leading measures, lagging outcomes, incident and near-miss reporting, feedback from users and affected people, and reassessment triggers.

A trigger converts evidence into management action.

If exception rates exceed a threshold, review may increase. If costs move beyond the approved range, the architecture may change. If quality falls for a particular group, the system may be restricted. If the intended value does not appear after the agreed period, the intervention may stop.

Without triggers, monitoring becomes observation without authority.

Minimum proof includes a measurement chain from technical performance to business outcome, full-cost visibility, monitoring after deployment, a cadence for reassessment, and defined events that cause scale, redesign, restriction, renegotiation, or termination.

5.14 Evidence must be able to change the intervention

An organization can satisfy the appearance of measurement while protecting the programme from its own evidence.

The dashboard is reviewed, but no threshold has authority. Incidents are logged, but the people reviewing the portfolio do not see them. Costs are reported after the contract renews. Employee feedback is collected, but the workflow owner is rewarded for adoption. Customer complaints sit in a different system from the AI performance report.

The result is evidence without consequence.

A working evidence system needs three connected loops.

The first is the learning loop before structural deployment. It tests the assumptions that matter most: task performance, user behaviour, integration, workflow effect, review burden, cost, or customer response. Its purpose is to reduce uncertainty, not to prove that the sponsor was right.

The second is the operating loop after deployment. It watches reliability, quality, exceptions, human intervention, incidents, cost, and changes in the environment. Its purpose is to keep the system within the conditions that management approved.

The third is the value and impact loop. It asks whether the intended outcome appeared, whether the benefit was realized rather than merely estimated, and where burden or harm accumulated. Its purpose is to decide whether the organization should expand, redesign, renegotiate, restrict, or stop.

The loops must meet in a decision forum with authority. Otherwise, the technical team can observe drift, Finance can observe cost, Operations can observe rework, and customers can experience failure while the portfolio remains green.

Learning is not the production of more information. It is the revision of an assumption, rule, resource allocation, workflow, or commitment in response to evidence.

■ KEY TERM

Three evidence loops

Learning before structural deployment, operating after deployment, and value and impact across both. They must meet in a forum with authority.

5.15 The weakest condition governs the decision

The framework's most important use is diagnosis.

When an initiative stalls, management often asks which team failed. The Five Conditions ask which condition is missing or contradictory.

Value clarity without workflow fit produces an important goal attached to an unusable intervention.

Workflow fit without operating rules produces an efficient path with uncontrolled permissions or inconsistent review.

Operating rules without ownership produce policies that no one has authority to resolve when trade-offs appear.

Ownership without evidence makes a person accountable for a story rather than an observable result.

Evidence without value clarity produces precise measures with no decision meaning.

The same problem can be misdiagnosed several ways.

Low adoption may reflect poor training. It may also reveal that the tool is badly placed, that employees are protecting quality, that incentives punish the new behaviour, or that the value claim was never meaningful to the user. Adding another training session treats the visible actor while leaving the system unchanged.

This is why the conditions should not be averaged too early.

A score of eighty per cent can sound ready. The missing twenty per cent may be the only person authorized to stop the system.

BOUNDARY

A score of eighty per cent can sound ready. The missing twenty per cent may be the only person authorized to stop the system.

Chapter 12 turns the conditions into a readiness judgment. Any scorecard used to record it must preserve minimum thresholds and non-negotiable blockers rather than letting strong dimensions compensate mathematically for an unusable one.

5.16 Proportionality without framework theatre

The Five Conditions can become bureaucracy if every intervention receives the same burden.

A private assistant helping one employee restructure non-sensitive notes does not require the same evidence, approval, and monitoring as an agent changing customer records or a model allocating access to a service.

The framework scales with consequence.

For a low-consequence, reversible experiment, value clarity may be a simple learning objective. Workflow fit may require only a task description and known data boundary.

Operating rules may fit on one page. Ownership may sit with one manager. Evidence may be a short test and explicit stop date.

As consequence, autonomy, volume, exposure, irreversibility, or stakeholder impact increases, each condition deepens.

The point is not to complete more paperwork. It is to make the hidden assumptions visible before they become structural.

The book's rule is simple.

The burden should match the decision being authorized.

5.17 The management standard

The Five Conditions are useful only when they alter the design or the decision.

A completed business case does not establish value clarity if the beneficiary and material threshold remain vague. A process map does not establish workflow fit if it omits exceptions and hidden review. A policy does not establish operating rules if people cannot tell what is allowed in the moment. A named sponsor does not establish ownership if the sponsor lacks authority over the changes required to realize the benefit. A dashboard does not establish evidence if no threshold can change the intervention.

The standard is coherence, not paperwork. The outcome, workflow, rules, owner, and evidence should describe the same version of the same use.

Enough is enough when management can name the uncertainty being accepted, the evidence required next, the people authorized to act, and the conditions that would stop the commitment from hardening further.

The Five Conditions are the minimum connected system. The first condition now becomes the next management decision: before selecting a use case, define the value case that would make the intervention worth pursuing.

KEY TAKEAWAYS

- The managed unit is one AI-enabled intervention inside real work – between enterprise strategy and technical components.
- Value clarity, workflow fit, operating rules, ownership, and evidence form a conjunctive test. The weakest condition governs.
- Proportionality prevents framework theatre: the burden matches the decision being authorized.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *Can we describe our largest AI use in one versioned record rather than five unrelated documents?*
- 02 *Which threshold, if crossed tomorrow, would change what this system may do – and who acts on it?*
- 03 *Which of the Five Conditions is weakest for our flagship use?*

06

CHAPTER 6 • PART II

Define the Value Case Before the Use Case

THE QUESTION

Why should this intervention exist, and what would prove it?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Outcome, beneficiary, and baseline	A visible value mechanism	Options, including no AI	Full cost and disbenefits	An owner and a next decision

DECISION RULE

Approve the uncertainty worth testing – not the technology the organization already wants to scale.

THE FIVE CONDITIONS • WHERE THIS CHAPTER SITS

Value clarity	Workflow fit	Operating rules	Ownership and accountability	Evidence and learning
---------------	--------------	-----------------	------------------------------	-----------------------

IN THIS CHAPTER

6.1 From value clarity to an investable hypothesis	71	6.6 The value case must include the model decision	76	6.11 Learning and option value are legitimate value cases	80
6.2 Start with the outcome and the comparison	73	6.7 Design the evidence before the environment changes	76	6.12 The value case is not the forecast	80
6.3 Make the value mechanism visible	74	6.8 Give the result an owner and a next decision	76	6.13 The burden should match the commitment	81
6.4 Compare options, including no AI	75	6.9 The value case is a contract between functions	77	6.14 The decision rule	81
6.5 Count the full cost and the displaced burden	75	6.10 Inside the first value-case review	78		

THE PHRASE “AI use case” often begins one decision too late.

Use AI to summarize customer calls.

Use an agent to answer billing questions.

Use machine learning to predict churn.

Use generative AI to draft proposals.

Each statement sounds practical because it names a capability and a place to apply it. None yet explains why the organization should invest, which outcome matters, who should benefit, what the current process costs, what would happen without the intervention, or what evidence would justify scale.

A use case describes what AI might do.

A value case explains why the intervention should exist.

The distinction is not academic. Once a use case enters a portfolio, the organization begins accumulating momentum around it. A sponsor appears. A vendor demonstrates a product. A technical team estimates feasibility. A pilot receives funding. Measures are chosen from whatever the tool can report. The project becomes easier to continue because people have already committed time, status, and budget.

The organization may eventually write a business case.

By then, the intervention is often inside the answer.

The stronger sequence begins with the change management wants to produce and compares credible ways of producing it. AI may be the preferred option. It may be one component of the preferred option. It may be useful only for learning. It may not be needed at all.

■ KEY TERM

Value case

A living explanation of an intervention's outcome, comparison, mechanism, alternatives, full cost, evidence, owner, and next decision.

6.1 From value clarity to an investable hypothesis

Chapter 5 established value clarity as one of the Five Conditions of AI Value.

Value clarity asks what worthwhile outcome should change, for whom, by how much, and by when. That is the minimum condition. This chapter turns that condition into a management instrument.

A value case is a living explanation of seven things:

the outcome the organization intends to change;

the baseline and counterfactual against which change will be judged;

the mechanism through which the intervention is expected to produce the outcome;

the alternatives management could pursue, including a non-AI option;

the full costs, disbenefits, risks, and dependencies;

the evidence and attribution logic;
and the owner, threshold, and next decision.
It is not a prediction dressed as certainty.

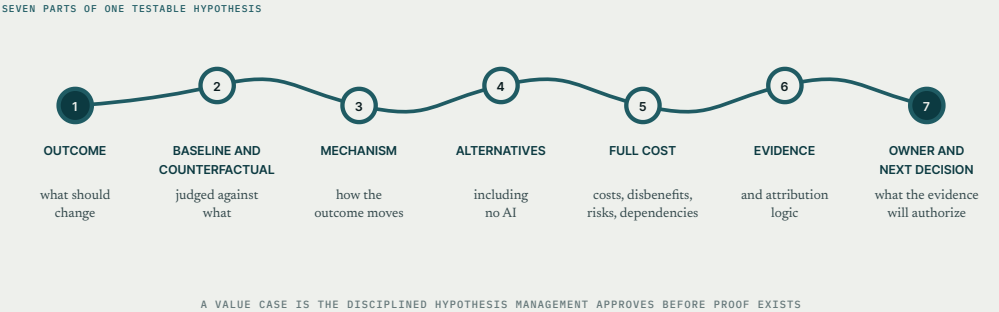
Early estimates will be ranges. Assumptions will be provisional. Some benefits will be non-financial. Some evidence will arrive only after a bounded test. The value case exists to make those uncertainties visible and governable before the organization mistakes enthusiasm for proof.

It also protects experimentation.
A small test should not need a thirty-page capital submission. But it should state what uncertainty it is reducing, why that uncertainty matters, how much the test may cost, and what evidence will determine the next move.

The principle is proportionality: the larger the cost, consequence, irreversibility, and organizational dependence, the stronger the value case must become.

The logic is not unique to AI. Serious appraisal begins with the objective, compares credible options, states the mechanism, establishes a baseline and counterfactual, counts full cost and disbenefits, and designs evidence before rollout removes the chance to collect it. AI makes that discipline more urgent because product momentum can form before the organization has defined the result.

FIGURE 6.1
The AI Value Case



The case connects outcome, alternatives, mechanism, economics, evidence, and the next decision.
SOURCE: AUTHOR.

The full value-case record lives in the companion toolkit. The narrative decision is simpler: management should be able to explain why the organization should act, how the intervention is expected to create value, and what evidence will authorize the next commitment.

6.2 Start with the outcome and the comparison

The value case begins with an outcome, not an AI verb.

“Use an agent for customer service” names a use case.

“Reduce repeat contacts for billing questions while maintaining resolution quality and making escalation no harder for customers” names an outcome with boundaries.

The outcome should be material enough to change a decision about capital, capacity, service, risk, or strategic direction.

The beneficiary must also be named.

Value can accrue to the enterprise, a business unit, employees, customers, partners, or the public. Those gains may not align. A service automation can lower company cost while increasing customer effort. A writing assistant can save one employee time while shifting review work to colleagues. A forecasting system can improve inventory turns while increasing supplier volatility.

A serious value case states whose outcome counts and which stakeholder effects could invalidate an apparently positive result.

This does not require every initiative to maximize every stakeholder’s benefit. It requires management to stop hiding distribution inside one aggregate number.

A baseline describes the current state before the intervention.

The counterfactual asks a different question:

What would probably happen without this intervention?

The distinction matters because the current state may be changing already. Demand may be rising. Staff may be learning. Another system may be launching. A policy may change. A seasonal effect may improve the metric. A declining business may produce lower cost because volume collapsed.

Comparing “after” with “before” is not always enough.

The organization needs the most credible comparison the decision warrants: a control group, staggered rollout, matched team, historical trend, process simulation, benchmark, or explicit business-as-usual scenario.

Perfect attribution is often impossible. That does not make comparison optional.

Without a baseline, improvement becomes memory.

Without a counterfactual, coincidence becomes value.

An initiative can improve a metric and still not deserve scale.

A one per cent reduction in handling time may be statistically real and economically immaterial. A quality gain may be valuable but arrive too slowly for the decision. A capacity benefit may matter only if the organization can redeploy or avoid cost. A revenue increase may be offset by acquisition, service, infrastructure, and risk costs.

The value case therefore needs a threshold.

DECISION

The threshold should match the decision, not the ambition.

Early experiments may use technical and behavioural thresholds because their purpose is to reduce uncertainty. Hard-to-reverse commitments need operating and economic thresholds because the organization is changing contracts, roles, customer promises, or capital allocation.

Time horizon matters too. An intervention may create immediate task gains but require several operating cycles before quality, retention, revenue, or avoided cost becomes visible. Delayed value is legitimate when milestones show how the capability is maturing. “Long term” is not a measurement plan.

6.3 Make the value mechanism visible

The value mechanism explains how the intervention is expected to move the outcome.

This is the bridge most use-case registers omit.

An assistant does not create value because it generates an answer. It may create value if the answer reduces search time, which allows more cases to be handled, which reduces backlog, which improves response time, which improves retention or avoids hiring.

Every arrow contains an assumption.

Will employees use the saved time for the intended work?

Will faster output preserve quality?

Will demand absorb the new capacity?

Will fewer routine tasks reduce staffing cost, or will the organization retain the same cost and gain resilience?

Will the AI reach the real bottleneck, or merely accelerate work before the bottleneck?

HM Treasury’s Magenta Book calls the broader discipline a theory of change: the causal chain from inputs through outputs to outcomes, together with assumptions, evidence, context, and possible unintended effects.¹

For enterprise AI, the mechanism can be concise. But it must exist.

If management cannot explain how the output becomes the outcome, it has a capability claim, not a value case.

IN PRACTICE

Draw the value mechanism as a chain of arrows. Under each arrow, write the assumption it depends on and the evidence that would test it.

6.4 Compare options, including no AI

A value case should not assume AI is the solution.

The organization may be able to change policy, simplify the process, improve search, remove duplicate approval, train employees, standardize data, purchase conventional automation, redesign the customer journey, or stop doing the work.

This step prevents the organization from paying an AI premium for a problem caused by poor process design.

It also improves technical decisions. Once management knows which outcome and constraints matter, it can compare a general-purpose assistant, retrieval system, predictive model, agent, conventional rules engine, or manual specialist process against the same value criteria.

The question is not “Can AI do this?”

It is “Which option produces the best result under the actual constraints?”

The option set should normally include business as usual, a do-minimum change, at least one non-AI intervention, one or more AI approaches, and a combined process-and-AI redesign where appropriate.

■ IN PRACTICE

Price five options side by side: business as usual, a do-minimum change, a non-AI intervention, an AI approach, and a combined process-and-AI redesign.

6.5 Count the full cost and the displaced burden

The visible price is rarely the full cost.

AI value cases should include software and model consumption, integration, data preparation, evaluation, security, governance, support, change management, human review, exception handling, monitoring, vendor management, and exit.

They should also include the cost of organizational attention.

A small licence can consume scarce engineering, legal, risk, process, and executive capacity. An initiative that looks cheap in procurement can be expensive in coordination.

Disbenefits belong in the same case.

A disbenefit is a negative consequence produced by the intervention: increased customer effort, slower exceptions, more surveillance, deskilling, concentration of vendor dependence, new error modes, environmental cost, or risk transferred to another team.

Counting disbenefits is not a moral add-on to the financial model. It prevents the sponsor from claiming value by moving cost outside the sponsor's ledger.

6.6 The value case must include the model decision

When model choice is material, the value case must compare the plausible operating configurations rather than assume the first workable model is the answer. The economic claim should identify a denominator close to the consequence – accepted output, completed case, protected decision, or realized outcome – and count the review, exception, remedy, integration, change, migration, and exit burden that sits around the model.

□ WATCH FOR

Value claimed by moving cost outside the sponsor's ledger – into review queues, customer effort, or another team's budget.

Chapter 9 owns the detailed model/system selection logic. The value case asks for the comparison and carries its result; it should not duplicate the selection method.

6.7 Design the evidence before the environment changes

The value case should specify the evidence before the intervention changes the environment.

Which data must be collected before launch?

Which groups, periods, or workflows will be compared?

Which leading indicators show that the mechanism is functioning?

Which outcome measures show that the organization is better?

Which technical measures are necessary but insufficient?

Who validates the result?

Not every intervention can support a randomized experiment.

A useful evidence design can still distinguish contribution from assertion. It can compare teams, model expected trends, trace process measures, investigate exceptions, and state what remains uncertain.

The objective is not to manufacture a perfect ROI number.

It is to produce evidence strong enough for the next decision.

6.8 Give the result an owner and a next decision

Every value case needs one accountable owner for the outcome.

That person does not own every specialist task. Finance may validate economics. Technology may own system performance. Risk may own control requirements. Operations may own the workflow. A product leader may own customer experience.

The accountable owner keeps the chain coherent.

The owner must have enough authority to commit the operational changes on which the benefit depends. A technology leader cannot realize labour savings that require a business unit to redesign roles. A project manager cannot improve customer retention without authority over the service. Finance cannot validate a capacity benefit that operations has not agreed to absorb.

The value case should also name who receives the gain and who carries the burden.

This makes a common failure visible: the sponsor owns the benefit while another function owns the work required to produce it.

The next decision

The final field prevents the value case from becoming a ceremonial approval document.

What decision will the evidence support?

Continue the experiment.

Change the workflow.

Switch the model or vendor.

Restrict the scope.

Renegotiate cost.

Deploy to a larger population.

Scale across the enterprise.

Stop.

The value case should be rewritten when the intervention changes materially. A different model, workflow, user group, autonomy level, or customer consequence can invalidate the original assumptions.

A value case that cannot change a decision is only a story about why the organization wants to proceed.

6.9 The value case is a contract between functions

A useful value case does more than justify funding. It creates a shared object around which functions can disagree productively.

Finance can challenge the materiality, time horizon, and full-cost assumptions. Operations can test whether the mechanism reaches the actual constraint. Technology can test feasibility, performance, and variable cost. Risk and legal can identify which disbenefits or consequences alter the threshold. Human resources can examine whether the proposed capacity benefit depends on role, incentive, or workforce changes. Customer and domain leaders can decide whether the stated outcome reflects lived value rather than internal convenience.

Without one coherent record, each function evaluates a different intervention. Finance reviews the original estimate. Technology changes the model. Operations changes the workflow. Risk approves a bounded user group. The sponsor later combines all four decisions into one success claim even though the approved assumptions no longer describe the system in production.

The value case should therefore carry a version. A material change to the model, data, workflow, population, autonomy, price, or consequence should trigger a review of the affected fields. Version control sounds administrative until an organization tries to explain why evidence from an earlier pilot still applies to a substantially different deployment.

When the baseline does not exist

Many organizations discover that the current process is not measured well enough to support a value case. Cost per case is unknown. Quality is judged informally. Exceptions are not recorded. Work moves through email and personal memory. The absence of a baseline is not a reason to invent one.

It is evidence about readiness.

Management has three choices. It can fund a short discovery period to observe the current process and build the baseline. It can use a bounded proxy while stating exactly what the proxy cannot prove. Or it can decide that the likely value is too small to justify the measurement investment.

■ IN PRACTICE

No baseline? Fund a short discovery period, use a bounded proxy with its limits stated, or conclude the value is too small to justify measurement.

The discovery work may itself produce value. Mapping the process can reveal duplicate approvals, avoidable demand, poor policy, or a conventional automation opportunity that is cheaper than AI. That outcome is not a failed AI project. It is the value case performing its first job: preventing the organization from solving the wrong problem expensively.

6.10 Inside the first value-case review

The billing-support assistant becomes useful only when the organization can make a decision about it.

The following scene is illustrative. It shows what the first value-case review should produce, not a universal meeting format.

The business owner circulates a two-page pre-read forty-eight hours before the meeting. It contains the current billing workflow, three months of baseline data, the draft value mechanism, five options, the expected cost range, known risks, and the decision requested. The meeting includes the business owner, contact-centre operations, a billing-

policy specialist, finance, technology, risk and privacy, a frontline supervisor, and the person responsible for customer complaints.

The sponsor opens with a familiar proposal: deploy a generative assistant to reduce agent workload.

The operations lead changes the question. The organization receives 120,000 billing contacts each quarter, and 31 per cent are repeat contacts within seven days. Difficult cases wait an average of 2.4 days for specialist review. The material problem is not generic workload. It is repeat demand, inconsistent first-contact resolution, and a growing specialist queue.

Finance asks what happens without the intervention. Contact volume has been rising, but a planned policy simplification may remove some avoidable demand. A before-and-after comparison would therefore overstate the assistant's contribution unless the rollout is staged across comparable teams.

The billing specialist identifies a deeper constraint. Agents do not struggle only because policy is difficult to search. Several policy pages conflict, and local workarounds are carrying knowledge that the official repository does not contain. A model grounded in the current source would generate consistent answers to an inconsistent policy.

Technology explains that a retrieval-and-drafting assistant is feasible, but the cost depends on volume, model choice, source maintenance, evaluation, logging, and how many cases require specialist review. Risk and privacy reject autonomous customer replies and account changes at this stage. The frontline supervisor warns that saving time on routine contacts could concentrate emotionally difficult and ambiguous cases among human agents.

The option set changes during the discussion. Business as usual remains the comparison. A do-minimum option improves policy search. A non-AI option simplifies the billing policy and expands specialist office hours. A rules-based option handles narrow, deterministic questions. The AI option retrieves approved policy and drafts a response for agent review. The combined option cleans the policy source first, then tests the retrieval-and-drafting assistant in selected teams.

Management selects the combined option.

The value hypothesis becomes specific: reduce repeat billing contacts by at least five percentage points within six months without worsening answer quality, complaint rate, adjustment error, escalation, or specialist backlog. The customer is the primary beneficiary through faster and more consistent resolution. Released agent capacity is a potential organizational benefit, not yet a booked saving.

The pilot will use a staggered rollout. Full cost includes licences, integration, policy maintenance, evaluation, agent review, support, and specialist workload. The customer-service executive owns the outcome. Finance validates the economics. Operations owns

the workflow. Technology owns system performance. The billing-policy owner is accountable for keeping the source current. Risk and privacy own mandatory boundaries.

The meeting does not approve enterprise deployment. It approves a bounded experiment with an explicit next decision.

The record that survives is more important than the meeting itself. It contains the outcome, baseline, counterfactual, options, mechanism, full cost, owner, evidence plan, permitted boundary, unresolved assumptions, thresholds, expiry date, and the decision management is authorized to make when the evidence arrives.

That is the difference between discussing a use case and governing a value case.

6.11 Learning and option value are legitimate value cases

Not every legitimate value case ends with immediate profit.

An experiment may build internal capability, reveal customer demand, test a vendor, reduce technical uncertainty, create a reusable data asset, or preserve a strategic option.

Those benefits should not be forced into false precision.

They should be named.

This protects innovation from two opposite errors.

The first is demanding mature financial return from an early exploration.

The second is calling every inconclusive pilot “valuable learning” after the fact.

A learning case should name the uncertainty, why reducing it matters, what evidence the experiment will produce, how much the organization is willing to pay for that evidence, and when the option expires or requires another commitment. Learning becomes value when it changes a decision.

6.12 The value case is not the forecast

Forecasts are useful because decisions concern the future. They become dangerous when the precision of the spreadsheet hides the uncertainty of the mechanism.

A projected return may depend on adoption, model performance, demand, labour redeployment, price, error rates, vendor terms, and the behaviour of several functions. The correct response is not to abandon forecasting. It is to expose the assumptions and show which ones the next experiment can test.

Ranges are often more honest than a single number. A scenario can show the minimum case, expected case, and upside case, with the variables that move the result. Sensitivity analysis can reveal that the investment is attractive only if an optimistic adoption rate or implausibly low review cost is achieved.

The value case should therefore separate what is known, what is estimated, and what must be learned. Management is not approving a promise. It is approving a sequence of increasingly informed decisions.

6.13 The burden should match the commitment

A value case can become bureaucracy when every experiment is forced through the same process.

HM Treasury's 2026 Green Book states the proportionality principle clearly: the effort used to develop a proposal should correlate with its scale, cost, complexity, and risk.²

The same rule belongs in AI management.

A reversible internal assistant used by five employees may need a one-page value hypothesis, a spend cap, basic data rules, and a review date.

A customer-facing agent changing accounts may need a detailed baseline, options appraisal, impact assessment, service and remedy design, security review, evaluation plan, and explicit authority to suspend.

The difference is not whether the project is called a pilot.

It is the consequence of being wrong and the cost of becoming dependent.

6.14 The decision rule

A value case is ready when it makes uncertainty visible rather than decorating a preferred solution. It is conditional when the outcome matters but the baseline, mechanism, comparison, or cost still requires a bounded test. It is risky when momentum around the tool shapes the assumptions. It is unusable when the document exists only to justify a decision already made.

Approve the uncertainty worth testing – not the technology the organization already wants to scale.

DECISION

Approve the uncertainty worth testing – not the technology the organization already wants to scale.

A value case is not proof.

It is the disciplined hypothesis management approves before proof exists.

The case names the outcome, comparison, mechanism, cost, evidence, owner, and decision. The intervention must still survive real work. Outputs may improve while

enterprise performance does not. Capacity may be created but never absorbed. Quality may rise for routine cases and fall for exceptions. Cost may move from labour to technology, review, or risk.

The next chapter follows that hypothesis through the evidence chain. A value case states what should happen. Operating proof shows whether it did.

KEY TAKEAWAYS

- A value case names the outcome, baseline and counterfactual, mechanism, options, full cost, evidence, owner, and next decision.
- Every value case compares against business as usual and at least one non-AI option.
- The case strengthens as cost, consequence, irreversibility, and dependence rise.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *Which of our current AI projects began as a use case without a value case?*
- 02 *What would probably happen without this intervention?*
- 03 *Who receives the gain, and who carries the burden?*

07

CHAPTER 7 • PART II

From Output to Outcome

THE QUESTION

Did the output change the outcome the organization cares about?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Technical output	Task effect	Workflow and the bottleneck	Stakeholder and financial consequence	Enterprise value

DECISION RULE

Evidence at one level authorizes only the next justified claim.

THE FIVE CONDITIONS • WHERE THIS CHAPTER SITS

Value clarity	Workflow fit	Operating rules	Ownership and accountability	Evidence and learning
---------------	--------------	-----------------	------------------------------	-----------------------

IN THIS CHAPTER

7.1	The missing conversion chain	84	7.6	The workflow decides whether capacity becomes useful	88	7.11	Four ways value claims break	91
7.2	What the evidence reveals across levels	85	7.7	The bottleneck decides	89	7.12	Delayed value still needs a path	94
7.3	Original research: operation has outrun attribution	85	7.8	Move from operating result to stakeholder, financial, and enterprise consequence	89	7.13	Inside the first measurement review	94
7.4	Technical output is necessary and insufficient	87	7.9	The denominator should follow the outcome	90	7.14	A positive outcome can still fail the value test	96
7.5	Task effect includes the human interaction	88	7.10	Enterprise value is a portfolio result	91	7.15	The measurement decision standard	96

CHAPTER 6 PRODUCED a value hypothesis. This chapter asks whether the evidence can carry that hypothesis through the organization.

In early 2026, a large international survey offered two statements that appear to conflict.¹ Across almost six thousand firms in the United States, United Kingdom, Germany, and Australia, 69 per cent reported using some form of artificial intelligence. Yet 89 per cent reported no labour-productivity effect over the previous three years. The same executives expected AI to increase productivity over the next three years.

The gap between observed past and expected future is not merely optimism waiting for time to pass.

It is a conversion problem.

Organizations can produce more AI output without changing organizational performance. Employees can write, search, summarize, and generate faster while throughput, quality, customer experience, cost, revenue, and return on capital remain unchanged.

That does not make the output useless. It means output and outcome are different objects.

BOUNDARY

Output and outcome are different objects.

The Microsoft 365 Copilot experiment introduced in Chapter 1 shows the divide at a smaller scale.² Workers who received access spent less time on email and reduced work outside regular hours, but the study did not establish average changes in the quantity or composition of work across the firms.

The intervention changed part of the workday. It did not prove that the organizations had changed their operating performance.

The management challenge is to connect the levels without pretending that one automatically proves the next.

7.1 The missing conversion chain

AI discussions often collapse six different claims into one sentence.

The model produced a good answer.

The employee completed the task faster.

The workflow improved.

The customer received a better service.

The company reduced cost or increased revenue.

The enterprise created durable value.

Each statement may be true. None is a synonym for the next.

A technically strong output can be ignored by users. A useful task gain can disappear inside an unchanged workflow. A faster workflow can produce more volume than the next stage can absorb. An operational improvement can create no financial benefit because capacity is not redeployed. A financial gain can be temporary because the cost, model, market, or customer response changes. A profitable intervention can still weaken resilience, trust, workforce capability, or strategic position.

The chain matters because organizations usually measure nearest to the technology.

Technical teams can see model quality, latency, and cost. Product teams can see active users and accepted suggestions. Employees can estimate time saved. Those measures arrive quickly and belong to identifiable owners.

Enterprise consequences arrive later and cross functions.

Operations must absorb the capacity. Finance must validate the economics. Customer teams must observe behaviour. Risk must track incidents. Human resources must understand changes in work. Strategy must decide whether the capability changes the company's position or merely keeps it at parity.

When the chain is not designed, every team can report success while the enterprise remains unchanged.

7.2 What the evidence reveals across levels

Current research shows that value appears unevenly across levels. Firms can report widespread use while seeing little firm-level productivity effect. Technical and adoption measures can improve before operating and financial consequences appear. Variable model, data, evaluation, orchestration, review, and support costs can make a cheap prompt expensive at the level of an accepted answer or completed business outcome.

The useful mechanism is not a contest between optimistic and pessimistic surveys. It is the chain that must carry a technical capability through human work, workflow performance, stakeholder consequence, economics, and enterprise decision.

7.3 Original research: operation has outrun attribution

The distinction between levels is visible in the 2026 Scalebotics AI Value Performance Benchmark.³ The research screened 200 companies and 300 public AI claim hypotheses. Two hundred and twenty claims from 164 companies contained enough public material for comparative evidence profiling across ten dimensions: boundary, workflow, outcome, baseline, economics, attribution, live evidence, governance, limitations, and traceability.

The strongest public evidence usually appeared near operation. Of the 220 public claims, 193 showed good evidence of real use. Only 47 showed a believable baseline or comparison, 36 meaningful evidence on cost, savings, revenue, or return, and 16 strong evidence linking the result to the AI. Governance was considerably more visible than counterevidence and organizational learning.

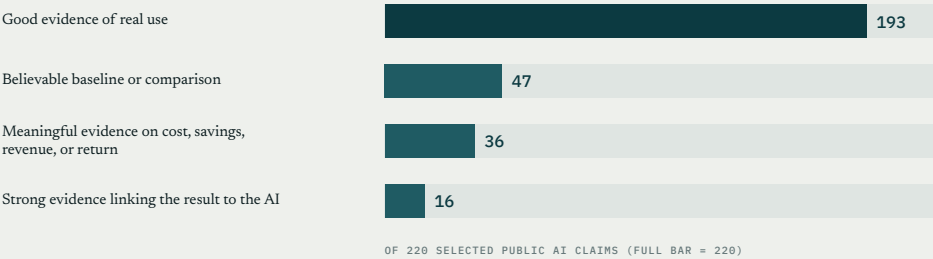
Those 220 claims come from the companies that disclose most. Beyond them, the record thins further. A wider 2026 sample screened 384 more claims; only 211 held enough public material to profile. Of those, 88 showed good evidence of real use and 3 showed strong evidence linking the result to the AI.

The benchmark measures what the public record authorizes an outside reader to believe about one selected claim. The same discipline applies inside the organization: management should not fill an evidence gap with narrative confidence simply because the intervention appears successful.

The pattern is useful because it shows where the proof chain becomes expensive. It is relatively easy to demonstrate that a system is live. It is harder to preserve a comparison while workflows change, harder again to expose full economics, and harder still to separate AI’s contribution from staffing, demand, policy, process redesign, and other simultaneous interventions. The Output-to-Outcome Chain is therefore not a demand for perfect causality. It is a way to keep the strength of the claim proportional to the evidence available.

EXHIBIT 7.1

What 220 public AI claims could show



Operation is far easier to evidence than comparison, economics, or attribution.

SOURCE: SCALEBOTICS RESEARCH, 2026 AI VALUE PERFORMANCE BENCHMARK (PURPOSIVE SAMPLE OF PUBLIC CLAIMS; METHOD IN NOTES).

The chain is causal, not ceremonial. The organization does not need perfect measurement before learning. It does need to know what kind of claim it is making and what evidence would support the next one.

7.4 Technical output is necessary and insufficient

The first level asks whether the system functions.

Does it produce an answer, recommendation, prediction, classification, draft, or action with acceptable accuracy, reliability, latency, security, and cost?

FIGURE 7.1

The Output-to-Outcome Chain



Evidence at one level authorizes only the next justified claim and commitment.

SOURCE: AUTHOR.

For a retrieval assistant, technical measures may include answer grounding, source coverage, latency, token cost, and failure rate. For a predictive model, they may include precision, recall, calibration, drift, and performance across relevant groups. For an agent, they may include tool success, action validity, recovery, and the rate of prohibited behaviour.

These measures are essential.

They do not establish that the output is useful inside work.

A model can score well on a benchmark while failing on the cases employees actually encounter. It can answer accurately but too slowly. It can be reliable in routine cases and dangerous in rare ones. It can perform well while costing more than the activity is worth.

Technical success authorizes a limited claim:

The system can produce under specified conditions.

It does not yet authorize deployment, value, or scale.

7.5 Task effect includes the human interaction

The second level asks what changes for the person or system performing the task.

Time may fall. Quality may rise. Errors may decline. A novice may gain access to stronger practice. An expert may produce more alternatives. Cognitive load may decrease – or increase because verification becomes harder.

The customer-support study by Brynjolfsson, Li, and Raymond provides a strong example.⁴ Access to a generative AI assistant increased issues resolved per hour by 14 per cent on average and by 34 per cent among novice and lower-skilled workers. The study also found improvements in customer sentiment and employee retention and suggestive evidence of learning.

This is more than output volume.

The assistant changed task performance and reached some adjacent outcomes inside one operating setting.

The heterogeneity matters. Experienced workers received little benefit. Averages can hide the population for whom the tool works, the population for whom it adds friction, and the expertise embedded in the system.

Task measures should therefore include more than speed.

The relevant question is not “How much did the model generate?”

It is “How did the task change after the full human interaction with the output?”

7.6 The workflow decides whether capacity becomes useful

Tasks create value through workflows.

A drafting assistant may reduce first-pass time while legal review remains the bottleneck. A coding assistant may increase code production while testing, security, and maintenance queues grow. Customer automation may remove routine contacts while complex cases wait longer. An underwriting model may speed assessment but create more appeals or manual exceptions.

Workflow performance asks whether the end-to-end process improved.

This level is where saved time either becomes useful capacity or disappears.

Suppose one thousand employees save one hour each week. The organization has created one thousand potential hours. The workflow realizes value only if those hours are absorbed into something the organization needs: more cases,

■ IN PRACTICE

Name the destination of released capacity before launch: more cases, higher quality, shorter queues, avoided hiring, or reduced overtime.

higher quality, shorter queues, customer work, learning, innovation, reduced overtime, avoided hiring, or resilience.

Workflow measures follow the completed service or process: cycle time, throughput, backlog, first-contact resolution, rework, defect escape, abandonment, exception volume, cost per completed case, and capacity at the constraint. Saving time before a bottleneck can increase congestion rather than value.

7.7 The bottleneck decides

Local productivity is valuable only to the extent that the workflow can use it.

If AI accelerates a non-bottleneck task, total throughput may remain unchanged. The organization produces more work-in-progress, more queueing, or more inventory before the constraint.

This is why process mapping matters.

The relevant unit is not the prompt, document, or employee. It is the completed service, resolved case, shipped product, approved decision, retained customer, or other end-to-end unit.

A strong value claim identifies where the constraint sits before deployment and tests whether it moves afterward.

When the bottleneck moves, the measurement system must move with it.

The tool may have succeeded at its original task while the organization must now invest somewhere else to realize the benefit.

7.8 Move from operating result to stakeholder, financial, and enterprise consequence

A workflow improvement matters because it changes an outcome for someone.

Customers may receive faster service, fewer errors, better access, or less effort. Employees may experience lower overload, stronger capability, or more surveillance and verification. Risk may fall through fewer incidents – or rise through a new failure mode. Regulators may receive better evidence. Partners may experience more stable demand or more volatility.

Business-unit outcomes may include retention, conversion, on-time delivery, compliance, service quality, inventory, loss, or resilience.

This level prevents the organization from treating internal efficiency as the whole result.

A service centre can reduce average handle time while repeat contacts and complaints increase. A sales team can produce more proposals while win rate falls. A fraud system can

reduce loss while declining legitimate customers. A generative search system can improve employee convenience while creating confidentiality exposure.

The outcome should match the beneficiary and boundary defined in the value case.

If employee sustainability is the declared outcome, reduced out-of-hours work can be value even when P&L does not change immediately.

If customer access is the declared outcome, the organization should not require every benefit to appear first as cost reduction.

Non-financial value is legitimate.

Unowned and unmeasured value is not.

Financial value appears when an operating outcome changes the economics of the organization.

Revenue increases.

Cost is removed or avoided.

Margin improves.

Working capital changes.

Loss, volatility, or expected risk falls.

Assets are used more effectively.

Capacity supports growth without proportional cost.

The full cost must be counted.

Licences, tokens, inference, cloud, data, integration, evaluation, support, monitoring, security, governance, vendor management, human review, incidents, and exit.

The useful unit is rarely cost per token.

7.9 The denominator should follow the outcome

Cost becomes decision-useful when its denominator follows the same conversion chain as the benefit. Cost per token or request can describe consumption; cost per accepted output or completed case is closer to the workflow; cost per realized outcome is closer to the consequence. The denominator should move only as far as the evidence can support.

Chapter 9 uses this same chain to compare configurations. This chapter keeps the accounting discipline: a technical or task gain is not enterprise value until the changed workflow, absorbed capacity, full cost, and stakeholder or financial consequence are visible.

■ KEY TERM

Denominator

Cost per token describes consumption. Cost per accepted output or completed case is closer to the work. Cost per realized outcome is closer to the consequence.

Meta’s Project OT illustrates the same category error at a different scale: a large increase in internal code-change activity did not translate proportionately into changes reaching users, while the reported operating burden rose elsewhere. Activity belongs on the chain. It should not be allowed to stand in for the outcome at the end of it.

7.10 Enterprise value is a portfolio result

The final level asks whether the intervention improves the enterprise after capital, risk, time, and alternatives are considered.

- Cash flow.
- Return on invested capital.
- Economic profit.
- Enterprise resilience.
- Strategic capability.

A single workflow may not move an enterprise metric visibly. That does not invalidate the workflow. The claim should remain at the level the evidence supports.

Enterprise value usually emerges from a portfolio of interventions, shared capabilities, and organizational learning rather than one isolated assistant.

This creates an aggregation problem.

Benefits can be double-counted when several tools claim the same hours. Shared infrastructure cost may be omitted from every project. One workflow may create data or capability used by another. Local cost reduction may be offset by central technology spending. Revenue gains may cannibalize existing business.

Portfolio value requires a common measurement architecture and a finance owner able to reconcile project claims with actual enterprise performance.

The organization should be able to move in both directions:

- From model performance upward to enterprise consequence.
 - From enterprise objectives downward to the workflows and systems expected to produce them.
- When those two views do not meet, the value story is incomplete.

□ WATCH FOR

Several tools claiming the same employee hours, and shared platform costs omitted from every project’s case.

7.11 Four ways value claims break

1. Potential capacity is reported as realized savings

Hours saved is attractive because it turns a complex change in work into one portable number. Ask employees how much faster they completed a task, multiply the answer by salary, and the organization appears to have produced a financial return.

The number can be useful. It is usually potential capacity, not realized value.

First, the estimate may be perceptual. Employees are good witnesses to their experience, but they may compare against an unusual task, forget verification time, or count minutes that would not otherwise have been paid or productive. Instrumented time data can be stronger, but it still does not reveal quality, effort, or what happened to the released time.

Second, an hour is not economically identical across contexts. Saving an hour at a bottleneck can increase throughput. Saving an hour in a task with spare capacity may change nothing downstream. Saving an hour for a scarce specialist can avoid delay or hiring. Saving an hour for an employee who is routinely working late may create workforce value by returning personal time rather than reducing cost.

Third, gross time saved can conceal new work. Employees review, correct, prompt, search for sources, manage exceptions, explain an output, or monitor a system. Another team may absorb the burden. The correct measure is net task and workflow effect, not the first visible reduction.

Fourth, capacity must be governed. Management must decide whether released time will support more volume, higher quality, innovation, training, resilience, reduced overtime, avoided hiring, or lower staffing. Without that decision, the benefit remains available but unrealized.

A credible time-savings claim therefore answers four questions: how was time measured, what new work was added, where the released capacity went, and which outcome changed because of that redeployment.

The organization should not discard hours saved. It should stop treating the hours as cash before anyone has changed the operating or financial system.

2. The review cadence ignores the time in the hypothesis

The conversion chain also operates on different clocks.

Technical measures arrive first. Adoption and task behaviour follow. Workflow measures need enough volume and time for queues, exceptions, and learning to stabilize. Customer, workforce, risk, and financial outcomes may appear over several operating cycles. Strategic capability can take longer still.

This timing creates two opposite errors. One organization declares success from early leading indicators. Another waits for a lagging financial result before learning whether the

mechanism is working.

Strong management uses both. Leading indicators test whether the conversion is progressing: the system is reliable, the intended users engage, accepted output rises, quality holds, the bottleneck moves, and exceptions remain manageable. Lagging indicators test whether the promised consequence arrived: customer behaviour changed, cost was removed or avoided, revenue or margin improved, risk fell, or a strategic capability became usable.

The review cadence should be explicit. Early reviews may occur weekly for reliability and task behaviour. Workflow and economics may be reviewed monthly or quarterly. The value case should state when an absent result becomes evidence against the intervention rather than merely evidence that more time is needed.

Time is part of the hypothesis. A benefit with no expected arrival date cannot be managed.

■ IN PRACTICE

State in advance when an absent result becomes evidence against the intervention rather than evidence that more time is needed.

3. Benefits are double counted while shared costs disappear

Value can also leak across organizational boundaries.

A business unit may claim the full benefit of faster work while central technology carries the model, platform, security, and support cost. Several use cases may claim the same employee hours. A revenue team may credit AI for demand that would have arrived through another channel. One intervention may reduce headcount while another counts the same released capacity as growth enablement.

The reverse can happen as well. A workflow may create reusable knowledge, data, evaluation tools, or controls that benefit later interventions, while the first project carries all of the cost.

Project economics and enterprise economics are therefore not identical. Finance needs rules for shared cost, benefit overlap, transfer, cannibalization, and reusable capability. Otherwise, a portfolio of individually positive business cases can add up to a disappointing enterprise result.

The final value claim should reconcile with actual budgets, operating measures, customer outcomes, and enterprise performance. If every project is winning but the company is not improving, the measurement system is counting the story more than the result.

4. Improvement is observed without a credible attribution design

Even when an outcome improves, the organization must decide how much credit belongs to the intervention.

Demand may change. Employees may learn. A policy may simplify. Staffing may increase. Seasonality may shift. Another system may launch. Managers may pay more attention because the workflow is being studied.

An after-versus-before comparison cannot separate those effects by itself.

Management should also preserve evidence of failure. Excluded cases, human overrides, customer appeals, and incidents are part of the outcome, not noise to remove from the dashboard.

The strongest feasible comparison should match the consequence of the decision: randomized or staggered rollout where possible, matched workflows, interrupted time series, process tracing, or another design that makes contribution more credible than assertion. The goal is not academic perfection. It is decision-grade evidence.

7.12 Delayed value still needs a path

Some AI value will take time.

Organizations must build data foundations, redesign workflows, develop skill, establish controls, integrate systems, and learn where the technology is useful. Earlier general-purpose technologies often produced delayed gains because complementary investments arrived before measured productivity.⁵

That history explains why immediate enterprise return is not the only valid standard.

It does not justify indefinite patience.

A delayed-value case needs leading indicators showing that the chain is moving: technical performance is stable, real workflow use is rising, task quality improves, the bottleneck moves, operating outcomes strengthen, unit economics approach the threshold, and the organization learns something that changes the next decision.

If the intervention repeatedly misses those gates, “the value will come later” becomes protection from evidence.

7.13 Inside the first measurement review

Six weeks into the illustrative billing-support pilot, the dashboard appears encouraging.

The assistant retrieves the approved policy source reliably for routine questions. Agents use the drafts frequently. Average handling time has fallen. The model team considers the pilot technically healthy.

The value-case review does not ask whether those facts are good news. It asks what they authorize.

The business owner, workflow manager, finance partner, technology lead, billing-policy owner, frontline supervisor, risk representative, and complaints lead meet with the

current record for the AI use. The numbers in this scene are illustrative.

At the technical-output level, source grounding and prohibited-content tests remain within the approved tolerance. Cost per draft is slightly higher than forecast because the system is using a more capable model for ambiguous questions.

At the task level, agents accept most drafts after editing, and handling time for routine billing contacts is down by 18 per cent. Verification effort is lower than the team feared. Technology argues that the assistant has passed the pilot.

Operations moves the discussion to the workflow level. Repeat contacts have fallen by only two percentage points, below the five-point threshold in the value case. Routine questions are faster, but ambiguous cases are reaching specialists more quickly and specialist backlog has grown by 22 per cent.

The complaints lead reports no broad increase in complaint rate, but customers with difficult adjustments are waiting longer for final resolution. The frontline supervisor says the tool is genuinely useful while warning that the team is concentrating harder work among fewer people.

Finance refuses to treat the 18 per cent handling-time reduction as a saving. Staffing has not changed, demand has not absorbed all of the released capacity, specialist overtime has risen, and policy-maintenance work is higher than forecast. The task benefit is real. The financial consequence is not yet positive.

■ KEY TERM

Realized benefit

A gain that has changed the operating or financial system – not potential capacity still waiting to be absorbed.

The chain makes the disagreement productive.

The technical team is correct that the output is reliable. The agents are correct that the task is easier. Operations is correct that the bottleneck has moved. Finance is correct that no cost benefit has been realized. The customer team is correct that the difficult-case experience has worsened.

No one needs to declare the pilot a success or a failure in the abstract.

Management decides not to scale. It narrows the assistant to routine billing questions, corrects the conflicting policy source, redesigns specialist routing, reserves capacity for the harder queue, and runs a second six-week comparison. The value case and record for the AI use are updated. The next review will decide whether repeat contacts fall enough to justify broader deployment.

The scene shows why the Output-to-Outcome Chain is a decision system rather than a dashboard. Positive evidence at one level can justify continuation while refusing a stronger claim.

7.14 **A positive outcome can still fail the value test**

Even a credible operating outcome may not justify continued investment.

An AI system can improve first-contact resolution, reduce defects, or increase conversion and still fail economically because the gain is too small, too expensive, too fragile, or inferior to another option. The comparison is not between the new system and doing nothing in the abstract. It is between the intervention and the best feasible use of the same capital, management attention, data, and technical capacity.

This is where materiality enters. A measurable improvement is not automatically a material improvement. Management should ask whether the change is large enough to alter a budget, customer decision, risk exposure, operating constraint, or strategic choice.

Durability matters too. A gain produced by exceptional pilot support may disappear when ordinary teams operate the system. A revenue lift may fade when competitors adopt the same capability. A cost benefit may vanish after vendor pricing changes or inference volume grows. A quality gain may deteriorate as the data, policy, customer mix, or model changes.

The value test therefore includes alternatives and time. Does the intervention outperform the non-AI option after full cost? Does the result survive broader use? Is the organization becoming more capable, or merely more dependent? Can the benefit be maintained without extraordinary attention?

A positive outcome authorizes celebration only at the level it has reached. Continued capital requires evidence that the outcome is material, economic, and durable enough for the next commitment.

DECISION

Continued capital requires evidence that the outcome is material, economic, and durable enough for the next commitment.

7.15 **The measurement decision standard**

Measurement is ready when it shows where the conversion succeeded, where it failed, and what management is authorized to do next. It is conditional when task or workflow gains are credible but materiality, economics, durability, or distribution remain uncertain. It is risky when the dashboard stops at the measures closest to the model. It is unusable when output, outcome, and enterprise value are presented as the same claim.

The output-to-outcome chain reveals a final fact.

AI does not create equal value everywhere in the workflow.

Its effect depends on where it is placed: before or after the bottleneck, inside or outside the decision, near or far from the customer, upstream or downstream of expert review, and at a level of autonomy the organization can absorb.

The next Part asks where AI belongs inside the task chain – because placement determines which work changes, where the bottleneck moves, and which authority and review design the organization will need.

KEY TAKEAWAYS

- Technical output, task effect, workflow performance, stakeholder consequence, financial value, and enterprise value are distinct claims.
- The bottleneck decides whether capacity becomes value. The cost denominator follows the outcome.
- Delayed value is legitimate only with leading indicators and a date on which absence counts as evidence.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *Where is the bottleneck today, and what evidence will show whether it moved?*
- 02 *Do our project-level benefit claims reconcile with actual budgets and enterprise performance?*
- 03 *Which of our reported savings are realized, and which are potential capacity?*



PART III

Designing AI into the Work

Part II established the value case and the evidence chain. Part III turns that hypothesis into an operating design: the exact place AI enters the work, the model configuration that belongs there, the authority attached to it, the review and escalation system around it, and the reusable governance that prevents every team from starting again from zero.

08	Where AI Belongs in the Workflow	100
09	Which Model Belongs in the Work?	109
10	Decision Rights, Human Judgment, and Escalation	117
11	Governance as Operating Infrastructure	128

08

CHAPTER 8 • PART III

Where AI Belongs in the Workflow

THE QUESTION

Where, inside the chain of work, should AI contribute?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Name the exact step	Start at the constraint	Separate friction from function	Test verification and reversibility	Design downstream absorption

DECISION RULE

A placement is credible only when both upstream and downstream are designed.

THE FIVE CONDITIONS • WHERE THIS CHAPTER SITS

Value clarity	Workflow fit	Operating rules	Ownership and accountability	Evidence and learning
---------------	--------------	-----------------	------------------------------	-----------------------

IN THIS CHAPTER

8.1	A use case is still too broad	101	8.5	Distinguish friction from function	104	8.9	Verification, reversibility, and consequence	106
8.2	Five places AI can enter the work	102	8.6	The capability frontier is local and temporary	105	8.10	Can the downstream workflow absorb the output?	107
8.3	Inside a workflow-placement session	102	8.7	Placement defines where routing may be useful	105	8.11	The placement decision	107
8.4	Start with the constraint, not the easiest demonstration	104	8.8	Context determines fit	106			

IN A FIELD EXPERIMENT with 758 management consultants, the same AI system produced two opposite results inside one profession.¹

For eighteen realistic tasks that fell within the technology’s capability frontier, people using GPT-4 completed more work, worked faster, and produced higher-quality answers than people working without AI. For a complex managerial problem deliberately chosen to sit outside that frontier, people using AI were less likely to reach the correct answer.

The workers were highly educated. The tasks resembled their ordinary assignments. They were free to ignore the system.

They still performed worse when AI was placed beside the wrong task.

The result is a warning against managing AI at the level of an occupation, department, or broad use case. The same workflow can contain one step where AI creates leverage and the next step where reliance on it damages judgment. The boundary is jagged, local, and temporary.

The management question is therefore not simply whether employees should have AI. It is where, inside a chain of work, AI should be allowed to contribute.

8.1 A use case is still too broad

“Use AI in customer service” does not define a placement. AI could classify an incoming request, retrieve account history, draft an answer, recommend a remedy, approve a credit, change an account, summarize the interaction, detect a complaint pattern, or monitor completed cases.

Those are different interventions. They operate at different points in the workflow, require different context, create different handoffs, expose different people to error, and demand different evidence and authority.

The same distinction applies in finance, law, health care, manufacturing, software, sales, and human resources. Saying that AI “supports underwriting,” “helps clinicians,” or “improves maintenance” hides the operating design.

Placement makes the design concrete. It identifies the exact step, the input, the output, the receiving actor, the next handoff, and the consequence if the system is wrong.

A valuable problem can contain a poor placement. A capable model can be inserted at the wrong stage. A task can improve while the bottleneck remains elsewhere. The value case and the Output-to-Outcome Chain therefore lead to a more precise question:

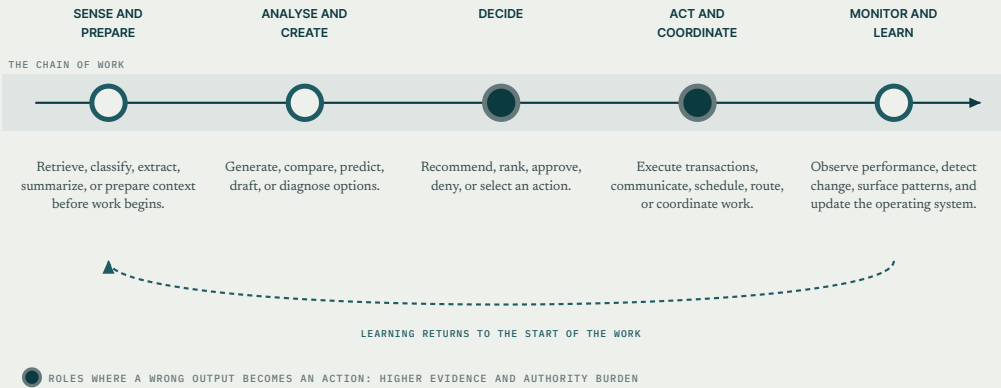
Where in the workflow can AI change the outcome without creating a larger problem around it?

■ IN PRACTICE

A placement record names six things: the exact step, the input, the output, the receiving actor, the next handoff, and the consequence if the system is wrong.

8.2 Five places AI can enter the work

FIGURE 8.1
Where AI Belongs in the Workflow



Five possible roles reveal the actual placement before authority is assigned.

SOURCE: AUTHOR.

These locations are not a rigid maturity ladder. One workflow may use AI in several roles, each with a different evidence burden. A retrieval component may remain approved while an action component is restricted. A monitoring system may continue after a drafting system is replaced.

The value of the model is not in naming five boxes. It is in forcing management to identify what the system is actually doing before technical permission becomes practical authority.

8.3 Inside a workflow-placement session

The following illustrative scene shows the decision a workflow-placement session should produce.² The format can vary.

A regional manufacturer operates six packaging lines. Unplanned stoppages have increased, and the maintenance backlog has become a board-level concern. The sponsor arrives with a clear proposal: deploy an AI maintenance agent that predicts failure and schedules work automatically.

The meeting includes the plant manager, the maintenance manager, a reliability engineer, a maintenance planner, an experienced technician, production scheduling,

finance, safety, and the team responsible for operational technology. They do not begin with the model. They begin by drawing the current chain of work.

A sensor or operator raises an alert. A shift supervisor decides whether the issue is urgent. A technician inspects the asset. The planner creates a work order, checks parts and labour, coordinates with production, and schedules a shutdown. The maintenance team performs the repair. Safety verifies isolation and return-to-service conditions. The reliability engineer later reviews failures and changes the maintenance plan.

The first surprise appears quickly. The plant does not mainly suffer from a shortage of alerts. It already receives more alerts than the team can examine. The deeper constraint is that information is fragmented across sensor dashboards, operator notes, maintenance history, manuals, parts records, and production schedules. Planners spend time reconstructing the case, while low-quality alerts compete with genuine deterioration.

The proposed placement – automatic scheduling – sits too late in the workflow. It assumes that the system already knows which alert matters, whether the sensor is trustworthy, what the current production condition is, whether the required part exists, and whether the asset can be isolated safely. Those assumptions are not yet true.

The technician points out another missing condition. Several machines behave differently when the plant changes packaging material or runs an unusual product mix. That context often lives in shift conversation rather than the maintenance database. A model trained on historical work orders may interpret an unfamiliar but acceptable pattern as failure, or treat a dangerous new pattern as ordinary.

The team separates what the system can do from what it should be allowed to do. AI can gather the alert, retrieve the relevant maintenance history and manual, identify missing information, draft a work order, suggest likely causes, and show whether the required parts are available. It can help the planner rank cases and later look for recurring patterns across completed repairs.

The system will not authorize a shutdown, change machine parameters, issue safety clearance, or return equipment to service. Those actions remain with people who can see the physical condition, understand production trade-offs, and carry the relevant authority.

The placement therefore moves earlier. AI will sit in sense and prepare, with a bounded role in analyse and create. The receiving actor is the planner, not the machine. The output must include the source records used, missing data, the reason for the priority recommendation, and the conditions that force direct inspection or escalation.

The workflow also changes downstream. The plant reserves a daily review window for high-priority cases, creates a route for unreliable-sensor alerts, and identifies who can resolve conflicts between production demand and maintenance risk. Without that capacity, better detection would simply create a larger queue.

The meeting ends with a placement record rather than a general endorsement of “predictive maintenance.” It names the exact tasks, users, information sources, prohibited actions, handoffs, verification method, fallback, and the evidence that would justify moving the system closer to scheduling or action later.

The model has not been rejected. The intervention has become more precise.

8.4 **Start with the constraint, not the easiest demonstration**

The easiest task to automate is not always the task worth changing.

A team can automate a report that nobody uses. It can generate more marketing content while approval remains the bottleneck. It can produce more software while testing and integration constrain release. It can classify claims faster while specialist review determines total cycle time.

Placement should begin with the value mechanism and the workflow constraint. What currently prevents the outcome from improving? Is the constraint information retrieval, expert availability, repeated drafting, coordination, approval, physical capacity, customer behaviour, or policy complexity?

AI belongs near the constraint only when it can alter it. If the bottleneck is a regulation, a scarce physical asset, a contractual dependency, or a decision management refuses to delegate, improving an upstream task may create more work-in-progress rather than value.

DECISION

Start with the constraint, not the easiest demonstration. AI belongs near the constraint only when it can alter it.

8.5 **Distinguish friction from function**

Organizations often insert AI into one step and preserve every surrounding handoff. The employee copies information into a tool, checks the response, reformats it, enters it into another system, seeks approval, and corrects the record after an exception. The model may be fast. The workflow is not.

Adjacent AI-suitable steps can sometimes be grouped so that repeated transfer, re-entry, and coordination disappear.³ Retrieval, source citation, drafting, and insertion into an employee’s workspace may create more value as one bounded chain than as four disconnected tools.

But handoffs also carry judgment. A specialist conversation may reveal tacit context. A review may detect that the customer’s situation is exceptional. An approval may resolve

conflicting objectives rather than merely add delay.

The placement task is to distinguish friction from function. Remove a handoff that only transfers information or repeats entry. Preserve or redesign a handoff that carries judgment, accountability, learning, trust, or remedy.

■ KEY TERM

Friction and function

A handoff that only transfers or re-enters information is friction. A handoff that carries judgment, accountability, learning, trust, or remedy is function.

8.6 The capability frontier is local and temporary

A benchmark cannot tell management where AI belongs in a specific workflow. The organization needs task-level evaluation using real inputs, ordinary users, edge cases, authoritative references, and the actual downstream recipient.

Performance depends on domain, data, prompt, model, interface, user skill, time pressure, and surrounding process. The frontier moves when the model, retrieval source, vendor, policy, or workflow changes.

A task that was outside the frontier can move inside it. A task that appeared reliable can move outside after a policy change or model update. A strong placement therefore includes a way to retest the boundary.

□ WATCH FOR

Placements approved once and never retested after a model update, a policy change, or a new workflow.

Broad rules such as “AI may draft but not decide” are useful starting points, not complete operating designs. Some drafts are consequential. Some decisions are routine and reversible. Some monitoring outputs quietly become decisions because employees treat them as authoritative.

The workflow has to be evaluated by effect, not label.

8.7 Placement defines where routing may be useful

Placement can identify where a route may be useful: ordinary cases may follow one path while difficult, ambiguous, or consequential cases escalate to a stronger system or human authority. This chapter establishes the task segments, volume, latency tolerance, context, consequence, reversibility, review capacity, and escalation needs that any route must respect.

8.8 Context determines fit

AI performs better when the information required for the task can be made available in a usable form. That information may include approved documents, current records, process state, customer history, domain rules, examples, and the reason an exception exists.

Not all context is stored. Experienced employees know which promise was made verbally, which supplier is unreliable, which policy is interpreted differently in practice, and which apparent anomaly is normal under local conditions.

A placement that depends on tacit context can create false confidence. The model receives a neat record while the real workflow depends on information outside it.

Management can capture and govern the missing context, narrow the AI role so that it does not require that context, or keep the step human-led. “More data” is not automatically the answer. The organization may lack the right to collect it, the context may change too quickly, or formalizing it may destroy discretion the work needs.

8.9 Verification, reversibility, and consequence

Two questions sharply change placement risk. Can a competent person or system detect a bad output before consequence? Can the consequence be reversed without disproportionate cost or harm?

A draft with visible sources is easier to verify than a persuasive recommendation whose basis cannot be inspected. A misrouted internal ticket can be corrected. A denied benefit, published statement, changed account, or unsafe physical action may be much harder to undo.

Easy verification and high reversibility support broader use. Hard verification and low reversibility argue for a narrower role, stronger evidence, or no AI placement at that step.

The closer AI sits to a customer, employee, patient, applicant, supplier, public decision, or physical process, the more its placement becomes part of the organization’s promise. Internal usefulness does not automatically authorize external exposure. External assistance does not automatically authorize autonomous action.

Physical workflows raise the threshold further. A maintenance recommendation still sits inside inspection, parts availability, production scheduling, safety isolation, and return-to-service authority. A wrong physical action can create scrap, injury, downtime, or environmental loss before feedback arrives.

■ IN PRACTICE

Ask two questions at every step: can a competent person detect a bad output before consequence, and can the consequence be undone?

Connectivity is not authority. The workflow must prove that an action is useful, bounded, observable, and recoverable under ordinary and exceptional conditions.

BOUNDARY

Connectivity is not authority.

8.10 Can the downstream workflow absorb the output?

An AI placement can succeed technically and fail because the next stage cannot use what it produces.

More leads can overwhelm sales. More detected anomalies can overwhelm investigators. More drafted contracts can overwhelm legal review. Faster triage can create a specialist queue. Better forecasting can be useless when procurement contracts cannot change.

The receiving stage needs capacity, incentives, compatible data, authority, and a reason to act. The workflow should therefore be mapped in both directions.

Upstream: what information and conditions does AI require?

Downstream: who receives the result, what do they do with it, and what new burden appears?

A placement is credible only when both sides are designed.

■ IN PRACTICE

Map downstream before deployment: who receives the output, what they do with it, and which new queue appears.

8.11 The placement decision

A strong placement decision may conclude that AI should assist at one step, recommend but not act at another, execute a tightly bounded action elsewhere, monitor the completed work, or remain outside the workflow under current conditions.

The decision should be versioned by component. Success in internal retrieval does not prove readiness for external advice. Success in drafting does not prove readiness for execution. Evidence transfers only when the task, context, user, consequence, and receiving stage are sufficiently similar.

Once management knows where AI belongs and what the work requires, another question becomes unavoidable.

Which model configuration is justified for this work, at this performance floor, consequence profile, and full economic burden?

Placement supplies the boundary. The next chapter chooses the model/system configuration that belongs inside it. Only after that choice is explicit can the organization allocate authority around the system it is actually proposing to use.

KEY TAKEAWAYS

- A placement names the exact step, input, output, receiving actor, handoff, and consequence.
- Start at the constraint. Remove friction; preserve function.
- Verification, reversibility, and downstream capacity decide how far AI may move toward action.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *Can a competent person detect a bad output before consequence – and can the consequence be undone?*
- 02 *Which of our placements depend on context that lives outside the system?*
- 03 *Who receives the output, and which new queue appears?*

09

CHAPTER 9 • PART III

Which Model Belongs in the Work?

THE QUESTION

Which configuration clears the performance floor at the strongest full economics?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Define the performance floor	Compare complete configurations	Follow both chains to value	Cost the unit nearest the consequence	Set the reopening triggers

DECISION RULE

Choose the configuration after the work is known – and let the decision expire.

THE FIVE CONDITIONS • WHERE THIS CHAPTER SITS

Value clarity	Workflow fit	Operating rules	Ownership and accountability	Evidence and learning
---------------	--------------	-----------------	------------------------------	-----------------------

IN THIS CHAPTER

9.1	The management question is not which model is best	110	9.5	Use the denominator nearest the consequence	112
9.2	Define the performance floor before choosing the model	111	9.6	Routing is an economic design choice, not a default	114
9.3	Compare complete operating configurations	111	9.7	The frontier moves, so the decision must expire	115
9.4	Two chains connect capability to value	112	9.8	The model-decision record	115

THE DEPARTMENT FOR Transport did not need the most powerful model in the abstract. It needed a system that could analyse consultation responses faster without losing the public voice that the process was supposed to hear.

Its Consultation Analysis Tool, co-developed with the Alan Turing Institute, used the Google Gemini family of large language models to generate and map themes across free-text responses.¹ The AI did not own the analysis. Human reviewers validated and refined themes before the results informed policy. Other models were tested during development, and model upgrades required prompts to be re-engineered and the pipeline to be evaluated again.

That detail matters because the model was only one component of the economic object. The published evaluation kept survey design, project management, human review, synthesis, and report writing inside the process instead of treating model output as the finished service. Across four principal live uses covering about 200,000 responses and more than eight million words, the Department estimated roughly £0.5 million in savings against rigorous manual analysis. It estimated that a medium consultation could cost 50 to 70 per cent less, while presenting broader annual figures as estimates rather than audited realized enterprise value.

This is the model decision management actually faces: which complete configuration clears the performance floor for this work and produces the strongest full economic result under the consequence and control burden the organization must carry.

9.1 The management question is not which model is best

Model markets invite a ranking reflex. One model leads a benchmark. Another is faster. Another has a larger context window. Another is cheaper per million tokens. Another can be deployed privately. Those facts matter, but none of them is the management decision.

The governing question is more demanding: Given a particular task, required performance floor, workflow and consequence profile, which model, combination of models or non-AI alternative produces the strongest full economic result?

The answer can legitimately be a frontier model. It can also be a smaller model, a retrieval system around a model, deterministic checks combined with a model, a route that sends different classes of work to different systems, a human-led process with AI assistance, conventional automation, or no intervention. The winner is local to the work.

This is why a benchmark leader can lose the business decision. A marginal capability gain may be irrelevant when the task is narrow. A cheaper endpoint may become expensive when it creates retries, specialist review, exceptions, failed actions, or customer remedy. A more expensive model may be cheaper per completed case when it materially reduces those burdens. A technically strong model may still be wrong when its

deployment, privacy, latency, switching, or concentration burden conflicts with the workflow.

9.2 Define the performance floor before choosing the model

Management cannot select economically until it knows what the work requires. The performance floor defines good enough for this task and this consequence. It should be strong enough to protect the outcome and narrow enough to avoid buying capability the workflow does not need.

DECISION

Define the performance floor before choosing the model.

Depending on the intervention, the floor can include task accuracy or quality, difficult-case performance, latency, throughput, structured-output reliability, tool or action reliability, context and retrieval behaviour, consistency across the target population, privacy or deployment requirements, inspectability, residual exception rate, and the amount of competent human review still required.

The performance floor is specific to the work and consequence, not a universal ranking. A customer-facing recommendation with difficult remedy may require a higher quality and review threshold than internal classification. A batch extraction process may tolerate seconds of latency that a live assistant cannot. A model that is excellent in English may fail the actual population.

A configuration that is acceptable for drafting may be unacceptable for execution.

■ KEY TERM

Performance floor

The quality, reliability, latency, and review burden this task and consequence require – strong enough to protect the outcome, narrow enough to avoid buying unneeded capability.

9.3 Compare complete operating configurations

A model is not deployed alone. The operating configuration includes the model family and version, deployment form, retrieval or context, prompts and tools, routing, deterministic checks, review, fallback, monitoring, and the workflow around them.

That wider object changes the option set. Management may compare a stronger model used broadly with a smaller model plus escalation. It may compare a general model with one grounded in authoritative retrieval. It may compare a single-model design with a route that reserves expensive reasoning for difficult cases. It may compare a cloud

endpoint with a controlled deployment, or an AI configuration with a conventional rules engine and human specialist.

The comparison should make the displaced burden visible. If a cheaper configuration adds an exception queue, that queue belongs to the option. If a retrieval layer requires continuous source governance, that work belongs to the option. If private deployment lowers data exposure but adds engineering and capacity cost, that belongs to the option. If a human-led alternative is slower but materially safer or easier to reverse, that belongs to the comparison.

■ IN PRACTICE

Assign every exception queue, source-maintenance task, and engineering cost to the configuration that creates it.

9.4 Two chains connect capability to value

The first chain follows value conversion:

model capability → task effect → workflow change → operating result → absorbed capacity → financial or stakeholder consequence

A model improvement creates value only if its effect travels through that chain. Faster drafting may create no benefit when review is the bottleneck. Better extraction may create no value when the downstream team cannot absorb more cases. Saved employee time becomes enterprise value only when the organization reduces cost, increases useful output, improves service, lowers loss, or uses the released capacity for something more valuable.

The second chain follows intervention design:

model capability → workflow → economics → controls → outcomes

Capability sets possibilities. Workflow determines how those possibilities enter real work. Economics tests whether the complete operating burden is worthwhile. Controls define what the system may do and what happens when conditions change. Outcomes decide whether the commitment deserves to continue.

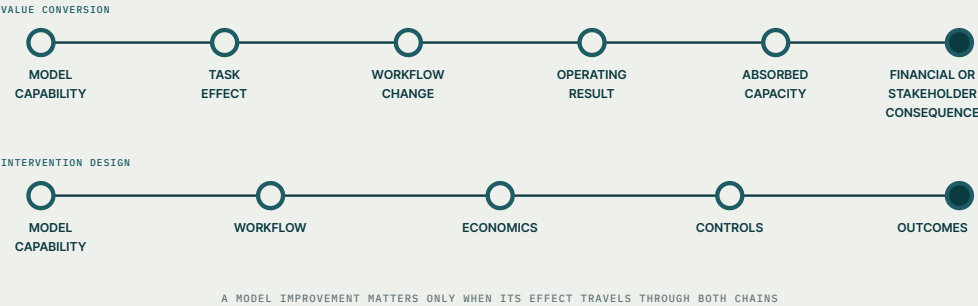
Their job is to keep the model from being evaluated in a vacuum.

9.5 Use the denominator nearest the consequence

For some workloads, cost per accepted output is the right unit: total operating cost divided by outputs that meet the declared standard. For a case workflow, cost per completed case is stronger because it includes routing, tools, retries, review, correction, exception handling, and recovery. For consequential applications, the meaningful denominator may be cost per protected decision or realized outcome. The closer the denominator sits to the consequence, the harder it is for hidden work to disappear.

FIGURE 9.1

Two chains connect capability to value



A model improvement matters only when its effect travels through the work, the economics, and the controls.

SOURCE: AUTHOR.

BOUNDARY

Token price is useful procurement information. It is rarely the management denominator.

Full economics should include the material parts of the operating system: model or licence consumption; retrieval, data, and external-service cost; engineering and integration; workload volume and peak demand; batching, caching, and asynchronous architecture where material; latency and capacity architecture; evaluation; human review and correction; retries; exceptions and remedy; monitoring; security and governance; vendor management; switching and migration; exit; and the destination of any capacity supposedly released.

The result can reverse an apparently obvious choice. A cheaper model can become the expensive option when the workflow spends heavily to make its output usable. A stronger model can justify a higher request cost when it reduces expensive failure modes. A small model can still dominate when the work is bounded, volume is high, and the quality floor is comfortably met.

CONTEXT • EVIDENCE

Morgan Stanley: the system around the model

Morgan Stanley makes this visible from a different angle.² Its wealth-management applications embedded GPT-4 inside a system of proprietary research, retrieval, evaluations, source links, advisor workflow, consent, CRM integration, and human

authority. The Assistant was built to help advisors retrieve internal knowledge; Debrief used Whisper and GPT-4 to turn consented meeting recordings into notes, action items, and draft follow-ups that advisors could review and send.

The public material reports very high adoption and a major expansion in the document corpus the Assistant could use. Their value for this chapter is architectural. The business object was not GPT-4 alone. The model became useful because the surrounding system supplied authoritative context, a real workflow, evaluation, controls, and a person who still owned the client relationship.

Another firm could buy the same model and create a worse system. It could also use a different model and create a better one.

9.6 Routing is an economic design choice, not a default

Model families increasingly make the trade-off explicit. OpenAI, for example, introduced GPT-5.4 mini and nano in March 2026 as faster, more efficient models for high-volume work, positioning nano for simpler tasks such as classification, extraction, ranking, and supporting subagents.³ Its own guidance makes the relevant point plainly: the best model for a workload is often not the largest one.

□ WATCH FOR

A cheaper endpoint that becomes the expensive option once retries, review, exceptions, and remedy are counted.

That is a product fact, not proof that a smaller model will create more business value. Management still has to test the actual task. A route can send ordinary cases to a cheaper or faster configuration while difficult or uncertain cases move to a stronger configuration and consequential cases move to competent human authority. But routing introduces classification error, monitoring, evaluation, fallback, and change burden of its own.

The economic question is therefore comparative. Does the route reduce total burden while holding the performance floor, or does it merely move cost into review and exceptions? Does the extra engineering complexity pay for itself at the expected volume? Can the organization observe the long tail separately rather than hide a weak tier inside an average quality score?

A one-model design can be the disciplined choice when volume is modest, the difference in unit economics is small, or routing complexity creates more risk than value. A multi-model design earns its place only when the evidence shows that the segmentation itself improves the intervention.

9.7 The frontier moves, so the decision must expire

Model selection is unusually perishable. New generations change the quality-cost frontier. Vendors change prices, rate limits, context windows, endpoints, deployment options, safety controls, and deprecation schedules. Retrieval and tool layers improve. The workflow itself learns.

Because the frontier moves, a materially changed configuration cannot simply inherit the evidence of the prior one. Model or prompt changes that can alter quality, burden, or economics require revalidation at the level they affect.

The model decision should therefore carry a change contract. Material changes to price, version, deprecation, latency, throughput, retrieval, tools, data route, deployment, vendor terms, or migration difficulty should reopen the evidence and economics at the level they affect. A current decision should not become a permanent vendor commitment by inertia.

■ IN PRACTICE

Write the reopening triggers into the record: price, version, deprecation, latency, retrieval, tools, data route, deployment, and vendor terms.

9.8 The model-decision record

Management does not need another branded canvas. It needs a decision record that can be inspected and reopened.

At minimum, the record should make nine things explicit: the required performance floor; the options compared; the selected configuration; the economic denominator; the major review and exception burden; the current evidence; material dependencies or concentration; the authorized boundary or next test; and the event that reopens the choice.

The toolkit carries those fields through the Value Case, Workflow Placement, Pilot, Runtime, and Portfolio records. The book keeps the logic visible: choose the configuration only after the work is known, count the full system rather than the endpoint, and keep the decision alive as the frontier moves.

Once the configuration is chosen, another question becomes unavoidable.

Who may frame, propose, approve, act, reverse, stop, investigate, and change the rules?

The model decision determines what system is entering the workflow. The next chapter allocates authority around that system and makes human judgment operational.

KEY TAKEAWAYS

- Define the performance floor before choosing a model. The winning configuration is local to the work.
- Compare complete operating configurations and count the burden each one displaces.
- Model decisions are perishable. Record the events that reopen them.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *Which denominator do we use – token, accepted output, completed case, or realized outcome?*
- 02 *Does routing reduce total burden at our actual volume?*
- 03 *Which vendor or version change would reopen our model decision?*

10

CHAPTER 10 • PART III

Decision Rights, Human Judgment, and Escalation

THE QUESTION

Who may frame, approve, act, reverse, and stop?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Placement creates power	Eight decision rights	Four review modes	Escalation is a timed transfer	Review capacity is economics

DECISION RULE

A review requirement with no capacity model is an unfunded control.

THE FIVE CONDITIONS • WHERE THIS CHAPTER SITS

Value clarity	Workflow fit	Operating rules	Ownership and accountability	Evidence and learning
---------------	--------------	-----------------	------------------------------	-----------------------

IN THIS CHAPTER

10.1 Placement creates power	118	10.5 Rights are real only when they can be exercised	120	10.10 Review capacity is part of the business case	124
10.2 The eight rights inside an AI-enabled decision	119	10.6 Human review has a job	121	10.11 The reviewer needs an independent line of sight	125
10.3 The right to frame is the first control	119	10.7 Review is not one mode	121	10.12 Near misses, complaints, and overrides form a sensor system	126
10.4 One accountable owner, several competent owners	120	10.8 Inside an authority and escalation decision	122	10.13 Authority and review must remain current	126
		10.9 Exceptions are evidence	123		

THE FACIAL-RECOGNITION SYSTEM did not physically confront customers. Rite Aid employees did.

From 2012 to 2020, the retailer used facial-recognition technology in hundreds of stores to identify people it considered likely shoplifters or other security risks. The system sent match alerts to store employees. According to a Federal Trade Commission complaint, many alerts were false positives.¹ Employees followed customers, searched or detained them, ordered them out of stores, publicly accused them, and called law enforcement.

The case is usually described as a failure of facial recognition. It was also a failure of authority design.

A probabilistic alert entered the workflow with enough practical force to trigger surveillance, exclusion, and police contact. Employees often lacked the confidence information behind the match. The organization failed to monitor false positives adequately and lacked a learning loop capable of converting repeated implausible alerts into a decision to restrict or stop the system.

The technology proposed. The organization acted. Accountability could not be assigned to the alert.

10.1 Placement creates power

Every AI placement allocates power, whether or not the org chart changes.

A retrieval system influences which facts become visible. A recommendation system influences which option receives attention. A ranking system determines which case is reviewed first. An agent with credentials can change records, commit funds, contact customers, or coordinate other systems. A monitoring system determines which patterns deserve investigation.

That is why decision rights cannot be reduced to the phrase “human in the loop.” A human can be present and still lack the time, information, skill, confidence, or permission required to change the result. A reviewer who sees a decision after the action is not an approver. An employee who can click override but is penalized for doing so does not possess a meaningful override right. A committee that approves a policy but cannot stop a system during an incident does not own suspension.

Decision rights are verbs. Who may define the problem? Who may propose an answer? Who may approve the action? Who may execute it? Who may reverse it? Who may stop the system? Who investigates what happened? Who may change the rules afterward?

■ IN PRACTICE

Name a holder for each verb: frame, propose, approve, execute, reverse, stop, investigate, and change the rules. A blank is a finding.

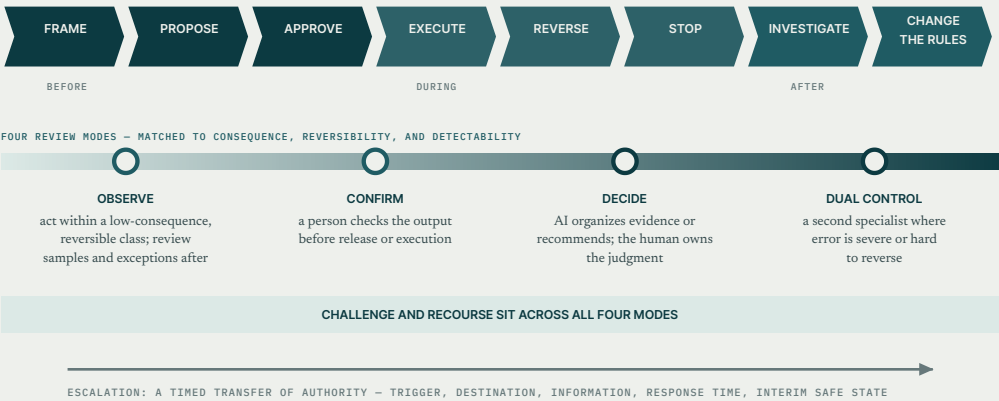
The answers determine whether AI is assistance, bounded delegation, or unmanaged authority.

10.2 The eight rights inside an AI-enabled decision

FIGURE 10.1

Decision Rights, Review, and Escalation

EIGHT DECISION RIGHTS – BEFORE, DURING, AND AFTER THE DECISION



Review activates rights and moves exceptions to a role with sufficient competence and authority.

SOURCE: AUTHOR.

10.3 The right to frame is the first control

Framing is often the least visible decision right because it happens before a formal decision appears. Someone decides what problem exists, which outcome matters, what data count as relevant, which alternatives deserve comparison, which stakeholders are in scope, and what would qualify as evidence. By the time a model is asked to recommend an action, much of the decision has already been made.

The right to frame should therefore be challengeable. Management should be able to ask who defined the problem, which proxy was selected, which alternatives disappeared before AI was considered, whose consequence is missing, and what new evidence would cause the organization to reopen the frame. A reviewer who may correct the answer but may not question the question is not exercising full judgment.

This matters especially when the AI output is persuasive. Fluency can anchor the discussion around the system's proposed interpretation. The stronger design gives at least

one competent person an independent standard of what good looks like and the authority to say that the problem has been posed at the wrong level.

This is a management right. A vendor may supply a model. A technical team may configure it. Neither should silently decide that cost is the correct proxy for medical need, conversion is the correct objective for a customer relationship, or a risk score is sufficient evidence to restrict an account.

As Chapter 4 showed, the reviewer must know what good looks like well enough to recognize when the system has answered the wrong question, used an unsuitable proxy, or operated at the wrong level of analysis.

10.4 **One accountable owner, several competent owners**

The business owner owns the intended result and the authority to change the workflow. Technology owns system reliability and access inside its competence. Data owners govern information quality and permission. Operations owns ordinary execution. Finance validates economics. Security owns incident containment. Risk, legal, privacy, safety, workforce, and domain leaders hold mandatory decisions where their expertise is required. Independent assurance challenges the evidence.

The accountable owner keeps those decisions coherent. A title-only owner is not enough. The person needs authority over the resources, process, incentives, and escalation required to realize the outcome.

Accountability cannot be outsourced to a vendor. A supplier may be responsible for service levels, security controls, defects, disclosures, and contractual remedies. It does not own the organization's customer promise, employment decision, safety duty, regulatory obligation, or strategic trade-off.

Committees can compare initiatives, establish policy, resolve conflict, and require evidence. They provide challenge. They do not substitute for ownership of the live workflow, customer case, or incident.

10.5 **Rights are real only when they can be exercised**

A documented right can still be ceremonial.

A reviewer may need a manager's approval that cannot arrive before the deadline. The interface may offer only approve or reject when the correct action is to request more evidence. Repeated disagreement may damage the reviewer's performance score. A temporary employee may be expected to challenge a system sponsored by senior leadership. A safety lead may possess a stop right on paper while lacking the technical permission to revoke the agent's credentials.

A real right requires information, competence, time, organizational protection, and an available action. The person must be able to pause the clock, transfer the case, request evidence, select a different outcome, record dissent, trigger remedy, or suspend the system within the scope of the right.

The decision-rights record should therefore name the role, scope, evidence required, transfer trigger, action available, and record that survives. It should also state what changes require renewed approval. Internal drafting does not automatically authorize customer communication. Recommendation does not automatically authorize execution. A material model or policy change can reopen the allocation even when the interface looks unchanged.

10.6 Human review has a job

Review is not made meaningful by placing a human somewhere in the workflow. The reviewer needs a defined purpose. In one intervention the job may be factual verification; in another it may be policy interpretation, professional judgment, consequence assessment, or a challenge to the frame itself. Different jobs require different evidence and different competence.

A reviewer can faithfully check whether an answer uses approved sources and still miss that the approved sources encode the wrong assumption. The system therefore needs to distinguish review inside a frame from review of the frame.

The purpose of review is not to reassure the organization that a person remains involved. Its purpose is to change the probability or consequence of a wrong decision.

A reviewer may prevent an incorrect output from reaching a customer. They may detect missing evidence, recognize that the case lies outside delegated scope, supply context the model cannot see, reverse an action, remedy harm, identify a recurring pattern, or suspend a system whose behaviour has changed.

A review step that cannot do any of those things may be theatre.

Meaningful review needs a trigger, a competent reviewer, independent evidence, sufficient time and capacity, practical authority, a durable record, and a safe state while the exception remains unresolved.

The label “human in the loop” guarantees none of them.

10.7 Review is not one mode

Different decisions require different review designs.

Observe means the system may act within a low-consequence, reversible class while people review samples, trends, complaints, and exceptions afterward. Confirm means a

person checks the output before release or execution. Decide means AI may organize evidence or recommend an option, but the human owns the judgment. Dual control requires a second specialist or independent reviewer where error is severe, conflict-prone, hard to reverse, or constrained by law or professional duty.

Challenge and recourse sit across all four modes. Affected people and frontline employees need a route to contest the result, supply missing evidence, request remedy, or signal a pattern management has not seen.

The appropriate mode depends on consequence, uncertainty, reversibility, detectability, volume, and available competence. Review everything is rarely the answer. Review nothing is not a design.

■ KEY TERM

Review modes

Observe, confirm, decide, and dual control. Match the mode to consequence, uncertainty, reversibility, detectability, and volume.

10.8 Inside an authority and escalation decision

The following scene continues the illustrative maintenance workflow from Chapter 8.

Before the assistant enters service, the manufacturer allocates the rights. The plant manager and maintenance manager frame the objective: reduce avoidable downtime without allowing production pressure to weaken safety or asset integrity. The AI may propose priority and likely causes. The planner may accept or change the work-order draft. The shift operations manager authorizes ordinary shutdown scheduling. A technician may take an asset to a safe state when an immediate hazard is suspected. Safety retains mandatory authority over isolation and return to service. The reliability engineer can change thresholds. The operational-technology incident lead can restrict or suspend the assistant. An independent safety and quality function can challenge the evidence and the record.

At 2:10 a.m., the assistant ranks a bearing alert as low priority and recommends deferral until the next planned maintenance window. The recommendation is plausible. Similar vibration patterns in the historical record did not lead to failure, the temperature remains within the ordinary limit, and stopping the line would disrupt a high-value production run.

The shift supervisor does not simply approve or reject the recommendation. The case was routed because two signals disagree. The interface shows the raw vibration trend, temperature, recent maintenance, sensor status, production condition, source records, and the limits of the model. An experienced technician notices that the line is running a heavier product configuration introduced after the historical data used for evaluation. The assistant is answering one question well – whether the signal resembles previous bearing

failures – while the operational decision requires another question: whether continued operation is acceptable under the current load and safety margin.

That is a framing exception, not merely an inaccurate output.

The supervisor transfers the case directly to the reliability engineer and operations manager. The line moves to a reduced safe state while they inspect the asset. The technician finds evidence of early degradation that is not yet visible in the work-order history. The team schedules an immediate controlled shutdown, replaces the component, and preserves the displaced production plan and customer commitments for operations to manage.

The recorded decision includes the original recommendation, missing context, who challenged the frame, the safe state used, the evidence inspected, the authority that changed the action, and the remedy. The assistant's role is narrowed until the new product configuration is represented in evaluation. Similar cases are rerouted for specialist confirmation. The value case is updated to include the additional review burden and the learning requirement.

No single participant “beat” the AI. The operating system allowed the organization to recognize that the system was answering from an incomplete frame and move the case to someone with enough competence and authority before consequence.

10.9 Exceptions are evidence

Exceptions reveal where the designed model of the work stops matching reality. A recurring exception may indicate a missing data field, a new customer pattern, an incentive conflict, a policy ambiguity, or a use boundary that was never as stable as the design assumed.

The goal is not to eliminate every exception. It is to route consequential exceptions safely and then ask which patterns should change the standard, the workflow, the model, or the value case. An exception system that only disposes of cases but never changes the intervention wastes one of its richest sources of operating evidence.

Some exceptions are expected scope exceptions: an amount exceeds the delegated limit, an image is too poor to interpret, or a requested action falls outside the ordinary process. Some reveal missing or conflicting information. Some reveal a system anomaly such as a spike in overrides, a vendor update, degraded performance in one language, or a security signal. Some reveal a policy conflict: the system behaves as configured, but the rule produces an outcome management no longer wishes to defend. Others are incidents requiring containment and remedy.

The full taxonomy and routing matrix belong in the operating toolkit. The narrative principle is simpler: route the case directly to the lowest role with enough competence,

authority, urgency, and independence to resolve it safely.

Escalation is a timed transfer of authority, not a vague instruction to “ask a manager.” It should name the trigger, destination, information that travels, response time, interim state, and action if the destination is unavailable.

DECISION

Escalation is a timed transfer of authority, not a vague instruction to “ask a manager.”

10.10 Review capacity is part of the business case

Capacity turns authority into economics. If ten per cent of cases require review and volume doubles, the review design is not the same intervention at a larger scale. It is a different operating system with different queues, staffing, response times, and failure modes.

A value case that counts automated throughput but omits reviewer minutes, specialist escalation, waiting time, and remedy work can make control look free. It is not. Review burden belongs in both the economics and the readiness decision.

A review requirement with no capacity model is an unfunded control.

BOUNDARY

A review requirement with no capacity model is an unfunded control.

Consider a workflow processing 100,000 cases each month. A five per cent review rate at four minutes per case creates roughly 333 reviewer hours before specialist escalation, remedy, or investigation is counted. If the exception rate rises to fifteen per cent and difficult cases take ten minutes, the burden becomes about 2,500 hours.

The model may remain technically efficient while the review design destroys the unit economics.

Management should therefore estimate expected volume, trigger rate, review time, specialist availability, queue behaviour, and what happens when demand exceeds capacity. The system should prioritize by consequence rather than merely arrival time. Service levels should make visible how long a case can wait before delay becomes part of the harm.

IN PRACTICE

Size review before launch: volume × trigger rate × minutes per case. Then rerun it at triple the exception rate.

Too much review can also make the system worse. Requiring approval for every low-consequence action can slow service, create queues, encourage rubber-stamping, obscure ownership, and push employees toward unofficial tools. The answer is proportionate review, not maximal review.

10.11 The reviewer needs an independent line of sight

Independence does not always mean organizational separation. It means the reviewer has enough information and permission to reach a conclusion that is not predetermined by the AI output or the sponsor's preferred story. That may require seeing original evidence before the recommendation, using a separate authoritative source, sampling cases the system did not flag, or receiving complaints and near misses that the model cannot observe.

When the only evidence available to the reviewer has already been selected and summarized by the system being reviewed, human oversight can become ceremonial.

A person cannot review what the interface conceals.

Many AI-assisted workflows present the recommendation as the dominant object and the evidence as a secondary link. The reviewer sees a fluent answer, a confidence label, or a ranked option before forming an independent view. That ordering can anchor attention and turn review into a search for reasons to approve what has already been proposed.

□ WATCH FOR
Interfaces that show the recommendation first and the evidence second. The ordering anchors the reviewer.

Meaningful review should expose enough of the case for the person to reconstruct the decision: the original request, authoritative sources, relevant policy version, missing fields, prior actions, model or rule version, uncertainty, and the reason the case was routed.

The interface should distinguish source evidence from generated interpretation. A citation that opens to the wrong passage is not evidence. A model-produced rationale is not the same thing as the record supporting the action.

Review design is therefore partly information architecture. Management decides what appears first, what can be inspected, what uncertainty is visible, which alternatives remain available, and how disagreement is recorded.

This is also where Engineered Epistemic Performance becomes a review problem. Reviewing whether an answer is acceptable inside the given frame is different from asking whether the frame – and the system's apparent competence inside it – is warranted for the decision. A reviewer should be able to ask: what evidence would make us conclude that the AI appears more authoritative or complete than the underlying evidence justifies? The question requires an independent standard, not a new committee or score.

10.12 Near misses, complaints, and overrides form a sensor system

Organizations often monitor confirmed errors and ignore near misses. Yet a near miss may be the earliest evidence that the system is reaching a dangerous boundary.

Overrides, abandoned drafts, repeated requests for clarification, manual rework, customer complaints, appeals, unusual response time, and specialist transfers are operating signals. None alone proves that the AI system failed. Together, they can reveal where the review design, data, policy, or workflow is degrading.

■ IN PRACTICE

Route overrides, complaints, appeals, and near misses to one owner. Separately, each looks like noise; together, they are the earliest warning.

Complaint and recourse channels are not only fairness mechanisms. They are sources of post-deployment evidence. A customer, employee, supplier, or frontline operator may observe a consequence the central dashboard cannot see.

Those signals must be connected. A complaint resolved in customer service, an override logged in operations, and a model anomaly recorded by technology may describe the same pattern. If each remains in a separate system, every function can report health while the intervention fails as a whole.

Review quality should be managed through latency, queue age, transfer and override patterns, remedy time, missed escalation, reviewer workload, repeated exceptions, and the later outcomes of approved and rejected cases. Very high and very low override rates are both ambiguous until management examines the cases.

10.13 Authority and review must remain current

The authority and review design is a living, versioned record. It should be reopened whenever the model, vendor, workflow, affected population, delegated scope, action class, or autonomy level changes materially.

The design is ready when the ordinary path is clear, exceptions can reach the right competence in time, reviewers can inspect independent evidence, intervention rights are practical, safe states and remedies exist, and operating signals can change the system.

Once those decisions have been made for one workflow, the enterprise faces a larger question: must every team rebuild the same rules, evidence, contracts, records, and specialist routes from the beginning?

The answer is governance as operating infrastructure.

KEY TAKEAWAYS

- Decision rights are verbs. Each needs a named holder who can actually exercise it.
- Meaningful review needs a trigger, competence, independent evidence, time, authority, a record, and a safe state.
- Review capacity belongs in the business case. Escalation is a timed transfer of authority.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *Could our reviewers pause the clock, request evidence, or suspend the system without personal risk?*
- 02 *Who holds the right to frame the problem, and can that frame be challenged?*
- 03 *What is our review burden if the exception rate triples?*

11

CHAPTER 11 • PART III

Governance as Operating Infrastructure

THE QUESTION

How does governance make good decisions cheaper to repeat?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Beyond the approval meeting	Standardize recurring decisions	Four pathways, not one gate	Governance as a service	The minimum durable record

DECISION RULE

Governance creates speed when the safe path is also the easy path.

THE FIVE CONDITIONS • WHERE THIS CHAPTER SITS

Value clarity	Workflow fit	Operating rules	Ownership and accountability	Evidence and learning
---------------	--------------	-----------------	------------------------------	-----------------------

IN THIS CHAPTER

11.1 Governance is not the approval meeting	129	11.5 Four pathways, not one gate	132	11.9 Assurance without transferred ownership	134
11.2 Two teams, two governance systems	130	11.6 Governance should be a service	133	11.10 Governance debt compounds quietly	135
11.3 Standardize recurring decisions, not every workflow	131	11.7 The minimum durable record	133	11.11 Governance can become anti-value	135
11.4 Governance as value infrastructure	131	11.8 Third parties remain inside the system	134	11.12 Governance does not prove readiness	136

MICROSOFT DID NOT try to govern every AI agent with the same rule.

Inside Microsoft Digital, employees could create simple information-retrieval agents with low-code tools.¹ More capable tools could connect to external channels, complete tasks, or automate workflows. Professional developers received access to the most powerful development environments. As capability and action increased, the governance burden increased with them.

The company describes governance embedded in creation and publishing tools, additional information-technology oversight where embedded controls were insufficient, and user education for risks no policy could anticipate completely. Agents were inventoried. Creation method, data access, sharing, audience, autonomy, and consequence triggered different policies and reviews.

This is company-reported practice from a vendor using its own products; its value here is the design pattern.

It still reveals the right distinction. Governance was not one committee standing between employees and AI. It was infrastructure built into how people discovered, created, registered, published, monitored, changed, and retired agents.

11.1 Governance is not the approval meeting

Organizations often encounter AI governance as a visible event. A team submits a use case. Risk, legal, security, privacy, technology, and the business ask questions. The proposal is approved, restricted, returned, or rejected.

The meeting matters. It is not the governance system.

BOUNDARY

The approval meeting is not the governance system.

The real system begins before the meeting and continues after it. Can the organization discover which AI capabilities are already in use? Does a routine internal assistant follow a different path from an agent that changes customer accounts? Are common evidence requirements already known? Can low-consequence work proceed under delegated authority? Does approval remain connected to the version actually operating? Are vendor changes, incidents, exceptions, and value results returned to the record?

When those mechanisms are absent, the committee becomes a substitute for infrastructure. Every team starts from zero. Every reviewer asks for different evidence. Similar uses receive different

□ WATCH FOR

Every team starting from zero, reviewers asking for different evidence, and similar uses receiving different decisions.

decisions. Business units learn to avoid the process. Specialist functions become overloaded. Shadow AI spreads while formal approval slows.

Governance then appears to be the source of delay, even though the deeper cause is that recurring decisions have not been made reusable.

11.2 Two teams, two governance systems

Consider two teams in the same enterprise. Both want to use an AI assistant to compare supplier contracts against the organization's approved clauses and identify provisions requiring specialist attention. The following contrast is illustrative.

In the first system, the team begins with a vendor demonstration. It then discovers that no one knows which contracts may be uploaded, which model version is permitted, whether the vendor retains the documents, what evidence legal requires, how security reviews the integration, who owns false negatives, which clauses require human confirmation, or where the use should be recorded. Each specialist asks for a different document. The team repeats explanations, changes the design several times, and eventually treats governance as the obstacle.

In the second system, the enterprise already has an approved retrieval environment for confidential documents, identity and access controls, standard vendor clauses, logging, a model-evaluation protocol, a data-classification rule, and a decision record that links the value case, workflow placement, authority, review, incidents, and version. A published pathway explains when internal drafting can proceed under delegated authority and when external communication or contract action requires legal authorization.

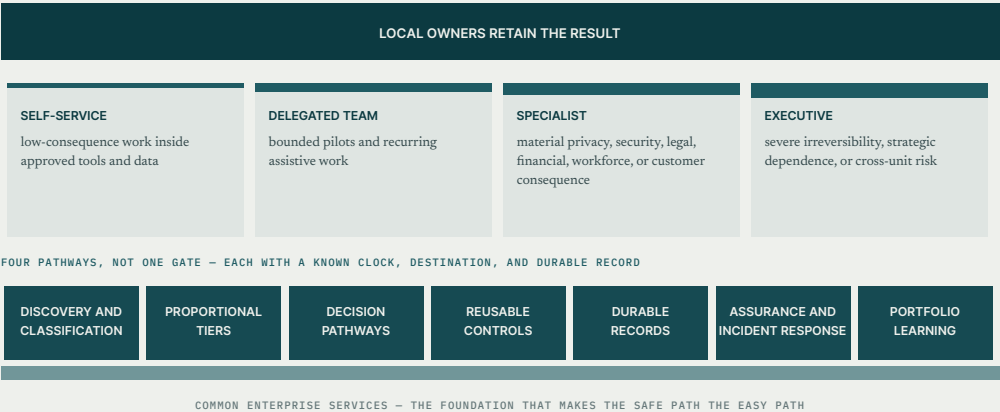
The reusable infrastructure does not approve the use automatically. It changes the quality and cost of the decision.

The business owner still has to explain why contract comparison matters, which workflow constraint it changes, what the full cost is, and whether the expected benefit is worth the review burden. Legal still determines which clauses require professional judgment. Procurement still owns the supplier relationship. Technology still validates the integration. Security and privacy still hold mandatory decisions inside their competence.

The governance service allows the team to start from an established operating surface rather than a blank page. The design is narrowed to internal comparison and drafting, source passages must remain visible, the assistant cannot amend or approve a contract, and uncertain or high-consequence clauses route to a lawyer. Governance has not merely passed or failed the proposal. It has improved the intervention.

11.3 Standardize recurring decisions, not every workflow

FIGURE 11.1
Governance as Operating Infrastructure



Common enterprise services reduce repeated uncertainty while local owners retain the result.

SOURCE: AUTHOR.

The enterprise needs one common infrastructure and many local owners. Central governance should not become the place where the business sends responsibility.

A committee can approve a customer-service agent. It does not own the customer relationship. A risk function can define mandatory controls. It does not own the service outcome. Internal audit can test the system. It should not operate it. A vendor can supply evidence. It cannot accept the organization's accountability to employees, customers, regulators, or the public.

This is a federated operating model, not a compromise between centralization and chaos.

11.4 Governance as value infrastructure

Infrastructure changes the economics of governance. A road does not decide where every driver should go. It establishes a common surface, signs, boundaries, and response systems that make many journeys possible. Enterprise governance should play a similar role.

It should standardize recurring decisions that otherwise produce confusion, duplication, or unowned consequence: discovery and classification; proportional tiers;

decision pathways; reusable controls; durable records; assurance and incident response; and portfolio learning.

This is how one intervention teaches the next one. A successful retrieval pattern can become an approved component. A repeated exception can change the standard. A vendor incident can alter contract requirements. A review bottleneck can change the evidence demanded before similar uses proceed.

Governance creates speed when the safe path is also the easy path.

DECISION

Standardize recurring decisions, not every workflow.

Governance creates speed when the safe path is also the easy path.

11.5 Four pathways, not one gate

A single approval route is easy to describe and difficult to operate. Proportional governance needs several decision paths.

A self-service path can cover low-consequence work inside approved tools and data boundaries, with clear prohibitions, logging, and a named process owner. A delegated team path can support bounded internal pilots and recurring assistive work using standard controls and evidence. A specialist path applies when the use creates material privacy, security, legal, financial, workforce, safety, or customer consequence. An executive or enterprise path is reserved for enterprise commitments with severe irreversibility, strategic dependence, large capital exposure, or risk crossing business units.

These are not permanent identities. A private assistant can move to a specialist path when it gains external data or action rights. A consequential system can move to a lighter renewal path after evidence, limits, and operating history reduce uncertainty.

Every route needs a known clock, a clear destination, and a durable record. Without a service level, proportional governance quietly becomes indefinite waiting.

The objective is not to push everything toward the lightest path. It is to place each decision on the least burdensome route that remains credible for the consequence involved.

■ KEY TERM

Four pathways

Self-service, delegated team, specialist, and executive. Place each decision on the least burdensome route that remains credible for its consequence.

■ IN PRACTICE

Publish a service level for every pathway. Without a clock, proportional governance becomes indefinite waiting.

11.6 Governance should be a service

The service test is practical: does governance reduce the amount of uncertainty every team must solve again? A useful enterprise service may provide an approved platform pattern, a standard evidence record, a privacy pathway, a vendor-control baseline, an incident route, a reusable model card, or access to a specialist who can make a bounded decision quickly.

The purpose is not to centralize every judgment. It is to make recurring decisions cheap enough that local teams can spend their attention on what is genuinely specific to the workflow and consequence.

Governance teams often describe what other teams must submit. Infrastructure asks what the governance system supplies.

It can supply approved environments, vendor intake, standard value and risk tiers, evaluation protocols, privacy and security patterns, reference architectures, human-review designs, contract clauses, monitoring requirements, incident routes, training, office hours, and reusable decision records.

A reusable control is more valuable than a repeated warning. An approved retrieval pattern can prevent every team from solving identity, source attribution, and access independently. A standard action boundary can make permissions explicit. A common record can connect value, authority, evidence, exceptions, and version. A vendor clause can preserve notice when a model changes. A central inventory can reveal duplicated spending and concentration risk.

The central function is not merely rejecting proposals. It is making acceptable designs easier to build.

11.7 The minimum durable record

The durable record is what prevents governance from existing only as memory. It should be possible months later to reconstruct what was authorized, under which boundary, using which evidence, with which restrictions, and what event was supposed to reopen the decision. Without that trace, organizations tend to inherit permissions after the reasons for those permissions have expired.

Policies govern categories. Organizations operate versions.

The governance record should connect the approved idea to the system currently running. At minimum, it should identify the intended outcome

■ IN PRACTICE

Link every approval to the version actually running – model, data, workflow, population, autonomy – and to the event that reopens it.

and beneficiary; accountable owner; workflow placement and affected population; model, vendor, data, integrations, and autonomy; applicable pathway; decision rights and review mode; evidence supporting authorization; approved version and date; material changes; incidents, complaints, overrides, remedies, and value results; and the next reassessment or retirement trigger.

The complete record belongs in the operating toolkit. The management principle belongs in the book: a record is valuable only when it can answer a live decision.

Can this use expand? Does a vendor update require retesting? Is the current deployment still inside the approved scope? Who can restrict it? What happened after the last exception? Are similar teams rebuilding the same control? Is the system still creating value?

A record that cannot answer those questions is an archive, not infrastructure.

11.8 Third parties remain inside the system

Most organizations will not own every model, tool, dataset, or hosting environment. Third-party dependence does not remove governance. It changes what must be governed.

A vendor can change the underlying model, retire a version, alter tool permissions or data handling, introduce a new agent capability, or raise inference prices without changing the interface employees see. The workflow can look unchanged while behaviour, exposure, and economics change materially.

Governance therefore needs notification rights, version visibility, evaluation triggers, fallback options, exit provisions, incident notice, and an owner who decides whether the prior authorization still holds.

The organization remains accountable for the use it chooses to operate even when another company controls part of the capability.

11.9 Assurance without transferred ownership

Independent assurance exists because the people building and operating a system face incentives, blind spots, and sunk costs. It should test whether evidence is credible, controls operate, records match reality, incidents are handled, changes are governed, and management claims remain supportable.

It should not become the de facto owner of the business result. The first line owns the use and its outcome. Specialist control functions define and challenge requirements inside their competence. Independent assurance tests the system and reports where management cannot rely on its claims.

Strong governance preserves both independence and local accountability.

11.10 Governance debt compounds quietly

When recurring decisions are not made reusable, each intervention carries the cost again. Teams re-negotiate data access, repeat vendor reviews, invent new approval language, rebuild monitoring, and search for the same specialist expertise. That is governance debt: unresolved enterprise questions being financed repeatedly through local project effort.

The opposite debt appears when an organization standardizes too aggressively. A control designed for one level of consequence becomes mandatory everywhere, low-risk learning slows, and people route work around the official path. Both forms create hidden cost. One duplicates uncertainty; the other duplicates burden.

Governance is infrastructure only when it makes good decisions easier without disconnecting them from local accountability.

11.11 Governance can become anti-value

Governance is not valuable by definition. It can become slow, defensive, repetitive, and politically useful to incumbents. Central functions can demand extensive documentation for trivial work. Committees can meet without authority. Specialists can issue contradictory requirements. Teams can wait while employees adopt public tools outside the system.

The remedy is not governance removal. It is governance design.

Time-box routine decisions. Publish delegated paths. Reuse evidence. Remove requirements that do not change a decision. Escalate only when consequence or uncertainty justifies it. Make disagreement visible. Preserve a route for experimentation. Measure whether the governance system itself is producing delay, duplicated control, shadow AI, or false assurance.

Useful operating measures include discovery coverage, time to decision, share of uses on standard pathways, evidence reused, specialist queue age, remediation time, repeated exceptions, decisions reversed, and whether the intervention's value and permission remain current.

■ KEY TERM

Governance measures

Discovery coverage, time to decision, share on standard pathways, evidence reused, specialist queue age, remediation time, and decisions reversed.

11.12 Governance does not prove readiness

Common infrastructure can make a sound AI use easier to build. It cannot prove that this specific team and workflow are ready.

An approved vendor does not establish a worthwhile value case. A standard review pattern does not guarantee enough reviewer capacity. A registered use does not show that the workflow can absorb the output. A governance pathway does not make poor data, unclear ownership, or weak evidence disappear.

Governance reduces repeated uncertainty. Readiness asks whether the required conditions actually hold here, for this use, at this level of consequence and commitment.

That is the next management decision.

KEY TAKEAWAYS

- Governance is operating infrastructure, not the approval meeting.
- Proportional governance runs on several pathways, each with a clock, a destination, and a durable record.
- Governance debt runs both ways: duplicated uncertainty and duplicated burden.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *Is our governance producing delay, duplicated control, shadow AI, or false assurance – and how would we know?*
- 02 *Does each approval link to the version actually running?*
- 03 *What does our governance function supply, not merely require?*

IV

PART IV

From Readiness to Runtime

Part III designed where AI enters the work, which configuration it uses, and who may frame, approve, challenge, restrict, and stop it. Part IV asks whether the people, workflow, data, and controls around the AI are ready for the next commitment, what a pilot must learn before that commitment grows, and how the organization keeps value and permission current after launch.

Readiness establishes what can be tried. A pilot establishes what has been learned. Runtime evidence determines what may continue.

12	Ready for What, Exactly?	138
13	Pilots That Produce Decisions	147
14	Keep Value and Permission Current	157

12

CHAPTER 12 • PART IV

Ready for What, Exactly?

THE QUESTION

What is this intervention ready to do next, for whom, and until when?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Declare the boundary	Judge the Five Conditions	Set critical gates first	Separate evidence from confidence	Contract the open conditions

DECISION RULE

The largest justified commitment follows the weakest unresolved critical gate.

THE FIVE CONDITIONS • WHERE THIS CHAPTER SITS

Value clarity	Workflow fit	Operating rules	Ownership and accountability	Evidence and learning
---------------	--------------	-----------------	------------------------------	-----------------------

IN THIS CHAPTER

12.1	Readiness is meaningless without a boundary	139	12.6	Inside a readiness decision	143
12.2	Not governance, maturity, or model quality	140	12.7	Test the organization around the AI before exposing the consequence	144
12.3	The Five Conditions become a readiness judgment	141	12.8	Conditional readiness is a management contract	144
12.4	Critical gates before averages	141	12.9	Avoid scoring theatre	145
12.5	Evidence grades and confidence are not the same thing	142	12.10	Readiness does not mean certainty	145

THE ALGORITHM COULD detect disease. The clinic could not always produce the input the algorithm required.

A Google Health research team studied a deep-learning system for diabetic-retinopathy screening across eleven clinics in Thailand.¹ The technical promise was substantial: analyse retinal images quickly enough to identify patients who needed specialist attention. The human-centred evaluation revealed a different constraint. Images were produced inside clinics with variable lighting, equipment, queues, local expertise, patient flows, and existing screening practices. The model had a data-quality threshold. The clinics did not always produce images that cleared it.

The study was designed to discover how the intervention behaved before wider deployment, and it did: model performance and system readiness are different claims.

A model can be ready in one sense while the organization is unready in another.

Deployment is where a technical result meets ordinary work: variable inputs, competing priorities, staff turnover, integration failures, queues, customers, exceptions, support costs, and consequences that a demonstration cannot reproduce completely.

Readiness is not a generic property of the organization or the model. It is a dated judgment about whether a declared system can support the next commitment.

12.1 Readiness is meaningless without a boundary

An organization may be ready to test an assistant with twenty trained employees and not ready to expose it to ten thousand customers. It may be ready for internal retrieval and not ready for external drafting. It may be ready for a recommendation and not ready for autonomous action. It may be ready at one site, on one channel, at one volume, and not elsewhere.

The readiness claim must therefore name the intended purpose, workflow placement, users and affected population, expected volume, geography and channel, data and systems, action and autonomy level, model and vendor version, fallback, and the period for which the evidence remains current.

Change the boundary and the decision may change.

■ IN PRACTICE

Declare the boundary before rating: purpose, placement, users, volume, channel, data, autonomy, model version, fallback, and expiry.

BOUNDARY

Readiness is readiness for the next commitment. Change the boundary and the decision may change.

This is the first protection against a familiar error: evidence gathered for a narrow experiment is used to authorize a broader operating system. A prototype may be ready for demonstration. A workflow may be ready for shadow use. A team may be ready for a bounded pilot. None of those statements proves readiness for production scale.

12.2 Not governance, maturity, or model quality

Governance can define the route, evidence burden, authority, and record. It cannot manufacture a missing condition. A committee can require human review. It cannot create reviewer capacity. A policy can require monitoring. It cannot make the decisive signal observable. An executive can approve a value case. Approval does not make the integration reliable, the data representative, or the fallback workable.

Enterprise maturity is also insufficient. A sophisticated organization can be unready for one intervention because its local workflow, owner, data, remedy, or economics are weak. A less mature organization can sometimes operate a narrow, reversible use safely because the boundary is clear, the consequence is modest, and recovery is easy.

The model is one component. A reliable model connected to stale policy is not ready. A strong retrieval system used by untrained employees is not ready. A capable agent without bounded permissions is not ready. A sound workflow with no capacity for exceptions is not ready. A technically successful use whose full operating burden exceeds the value it creates is not ready.

The critical dependency may be technical, operational, financial, organizational, or consequential. That is why readiness cannot be averaged casually.

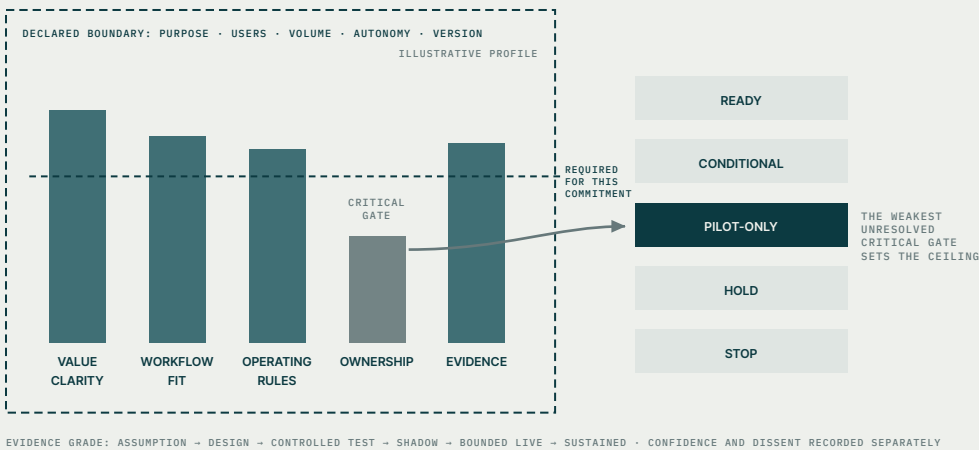
Technical readiness should therefore be declared at the same level of specificity as the operating boundary. Where material, the record should show task-specific evaluation, expected throughput and rate limits, latency tolerance, the exact model version or snapshot, fallback, deployment and data-route conditions, and the privacy or security constraints that make the proposed configuration usable. A generic statement that the model performs well is not enough.

Readiness also has an economic ceiling. The question is not only whether the system can operate, but whether the evidence supports this scale, volume, review burden, vendor dependence, and cost. A technically successful configuration can be ready for a shadow test and still be unready for a commitment whose unit economics have not survived ordinary workload and exception conditions.

12.3 **The Five Conditions become a readiness judgment**

The Five Conditions remain the organizing system. Readiness asks whether each condition is strong enough for the boundary and consequence now being proposed.

FIGURE 12.1
Readiness for the Next Commitment



The boundary, Five Conditions, critical gates, evidence, and next justified commitment remain connected.

SOURCE: AUTHOR.

A readiness scorecard records this judgment: the boundary, target, evidence, criticality, confidence, dissent, and next decision in one place.

I do not regard a score as a substitute for judgment. A useful scorecard exposes where judgment is being exercised. A bad one hides judgment behind arithmetic.

12.4 **Critical gates before averages**

A readiness decision is about the next commitment, not the organization's general virtue. That is why critical gates matter more than a flattering composite. If the proposed use can affect a customer's money, a missing remedy path may be decisive even when data, integration, and model performance are strong. If the use is a reversible internal assistant, the same gap may be containable through a narrower boundary.

The gate is therefore not a universal checklist item. It is the condition that cannot remain below the required level for this consequence and this commitment.

A readiness instrument becomes dangerous when a favourable average conceals one condition that can invalidate the deployment. Identity failure, inability to reverse a harmful action, absent legal authority, unreliable source data, no remedy, or no fallback may be decisive even when the rest of the profile looks strong.

Set criticality before scoring. Otherwise, teams tend to declare the weakest condition non-critical after the result becomes inconvenient.

□ WATCH FOR
Teams declaring the weakest condition non-critical after the score becomes inconvenient.

The largest commitment the evidence can justify follows the weakest unresolved critical gate.

A use may be ready for technical testing but not live data. Ready for shadow operation but not customer exposure. Ready for recommendation but not action. Ready for one site but not scale.

The ceiling is not punishment. It is the maximum commitment the current evidence can honestly support.

12.5 Evidence grades and confidence are not the same thing

Readiness also fails when evidence strength and reviewer confidence are blended into one impression. A team can be highly confident about a design that has never operated. Another team can hold strong live evidence but disagree about what it means. Those are different situations and they should produce different next decisions.

Evidence grade describes what kind of observation exists at the defined boundary: assumption, documented design, controlled evaluation, shadow operation, bounded live evidence, or sustained operation. Confidence describes how much uncertainty remains in the interpretation. Dissent records where competent people reach different conclusions. Keeping the three visible prevents a confident sponsor from converting thin evidence into apparent maturity.

■ KEY TERM
Evidence grade
Assumption, documented design, controlled evaluation, shadow operation, bounded live evidence, or sustained operation. Confidence and dissent are recorded separately.

The evidence must remain current as well. A strong result from an older model version, lower volume, different population, or different workflow may no longer support the same authorization. Readiness has an expiry because the intervention has a boundary.

Dissent also belongs in the record. When operations, engineering, finance, legal, or frontline staff interpret the same evidence differently, the disagreement should not be averaged away. It should be resolved by authority, narrowed through an experiment, or preserved as an explicit condition of the decision.

12.6 Inside a readiness decision

The following scene is illustrative. It shows how a readiness judgment should constrain a commitment; it does not describe a specific bank.²

A regional bank has tested an AI system that ranks card transactions by fraud risk. Fraud analysts currently review alerts and decide whether to contact the customer, allow the transaction, or place a temporary hold. The sponsor proposes a broader use: the system should automatically place short holds on transactions above a high-confidence threshold, with analysts reviewing the remaining cases.

The readiness meeting includes the payments executive, fraud operations, customer service, complaints and remedy, privacy, model risk, engineering, finance, and an experienced analyst. The technical team presents a strong retrospective result and a successful six-week shadow test. The system identifies more confirmed fraud than the existing rules at the same alert volume.

The first disagreement concerns the target. Engineering rates performance as ready because precision cleared the agreed threshold. Customer service rates workflow fit below target. Even a small false-positive rate at the proposed volume would create several hundred additional calls during peak periods, and the after-hours team cannot release a hold without escalation.

The complaints lead raises a deeper issue. The shadow test recorded whether an alert was correct, but it did not test how quickly a legitimate customer could regain access to funds, whether the explanation was understandable, or whether the burden fell unevenly on travellers, shift workers, or small businesses making unusual purchases. The value case counted prevented fraud and analyst capacity. It did not count customer remedy.

Model risk asks whether the missing evidence is merely desirable or critical. The group classifies the ability to reverse a false hold promptly as a critical gate. The intervention changes access to money. A high-performing model cannot compensate for a remedy process that has not been tested.

The evidence grades now matter. Model performance has controlled-test evidence. Analyst workflow has shadow-operation evidence. Customer remedy, peak-volume capacity, and segment effects remain design assumptions. The overall average would still look favourable. The critical gate sets a lower ceiling.

The decision is not “approved” or “rejected.” The bank authorizes a narrower commitment. The system may continue ranking alerts and may prepare a recommended action, but it may not place a hold autonomously. A thirty-day pilot must test peak call volume, release times, customer communication, segment-level false positives, after-hours authority, and the manual fallback. Authorization expires unless the evidence is reviewed.

Any material complaint pattern, queue beyond declared capacity, or inability to release a hold within the stated service level returns the system to ranking only.

The meeting ends with owners, due dates, evidence required, and the next decision. The model has not become less capable. Management has become more precise about what capability may authorize.

Readiness evidence must also survive the operating handoff. A project team may understand every caveat while the receiving manager sees only a green status. The decision record should carry the unresolved assumptions, temporary controls, evidence expiry, and the authority needed to reopen the intervention into ordinary operation.

12.7 Test the organization around the AI before exposing the consequence

Readiness includes more than the model and integration. It includes the people who will notice failure, the queues that receive exceptions, the data that must stay current, the manager who can suspend the use, the customer or employee who needs remedy, and the budget that absorbs review and support.

This is why exceptional pilot teams can create false readiness. They remember caveats, work around broken interfaces, know which specialist to call, and compensate for missing documentation. Before exposure increases, management should test whether the ordinary organization – normal staffing, normal workload, normal handoffs – can carry the intervention without relying on heroics.

■ IN PRACTICE

Test with normal staffing, normal workload, and normal handoffs. Remove the developer from daily rescue and observe what happens.

Readiness evidence should come from the environment that will receive the intervention. Silent tests, shadow operation, parallel runs, simulations, red-team exercises, and limited populations can reveal data-pipeline instability, workflow burden, local performance, exception volume, integration fragility, and recovery limits before the system controls a live consequence.

Readiness therefore asks which protective work made the test succeed, whether it remains necessary, who will perform it in operation, and what it costs.

12.8 Conditional readiness is a management contract

Many interventions will be neither fully ready nor unusable. Conditional readiness is valuable only when the condition is operational.

A condition must name the missing state, the accountable owner, the evidence required, the allowed boundary while the gap remains, the deadline, the trigger that

narrows or stops the intervention, and the decision that follows closure.

Planned training is not readiness. Promised monitoring is not readiness. An integration on the roadmap is not readiness. They are obligations until they exist and have survived a test.

□ WATCH FOR

Readiness built on promises: training planned, monitoring promised, integration on the roadmap.

Temporary controls also need expiry. A manual review that quietly becomes permanent can reverse the economics. A specialist queue that never receives capacity can turn governance into delay. A workaround that depends on one employee can become hidden operational debt.

12.9 Avoid scoring theatre

A scorecard can improve consistency. It can also create false precision, encourage teams to optimize the form, and turn a contextual judgment into a portfolio beauty contest.

The prevention is simple but demanding. Define the boundary and target before rating. Rate the condition, not the team's effort. Attach evidence. Identify criticality first. Keep confidence separate. Preserve dissent. Name the scorecard owner. Calibrate reviewers against the same cases. Do not compare unlike interventions as though a low-consequence drafting assistant and a system affecting access to money belong on one universal scale.

The output should be a decision profile, not a decorative number: ready, conditional, pilot-only, hold, or stop – together with the evidence and obligations that make the classification credible.

12.10 Readiness does not mean certainty

No organization can prove every condition before real use. Demanding certainty would freeze learning and push experimentation into informal channels. Some evidence can be produced only in operation.

The answer is not to abandon readiness. It is to reduce the commitment. Limit the population. Remove external communication. Disable action. Use shadow mode. Shorten the authorization period. Increase sampling. Add independent review. Fund the missing capacity. Create a stop trigger.

Readiness is not perfection. It is a credible match between evidence, boundary, consequence, and the organization's ability to recover.

DECISION

Do not ask whether the organization is ready for AI. Ask what this intervention is ready to do next, for whom, under which conditions, and until when.

KEY TAKEAWAYS

- Readiness is a dated judgment about a declared boundary, not a property of the organization or the model.
- Set critical gates before scoring. The weakest unresolved gate caps the commitment.
- Keep evidence grade, confidence, and dissent separate. Open conditions must be operational contracts.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *Which of our readiness conditions are still promises?*
- 02 *Would this use survive normal staffing, normal workload, and normal handoffs?*
- 03 *When does our current readiness evidence expire?*

13

CHAPTER 13 • PART IV

Pilots That Produce Decisions

THE QUESTION

What decision will this pilot's evidence authorize?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Start from the decision	Name the primary uncertainty	Write the failure conditions	Test the ordinary system	End in a decision meeting

DECISION RULE

A pilot succeeds when it changes what management is justified in doing.

THE FIVE CONDITIONS • WHERE THIS CHAPTER SITS

Value clarity	Workflow fit	Operating rules	Ownership and accountability	Evidence and learning
---------------	--------------	-----------------	------------------------------	-----------------------

IN THIS CHAPTER

13.1	Technical test, proof of concept, pilot, and deployment	148	13.5	The Managed Pilot Contract	150	13.10	Evaluation needs enough independence	154
13.2	Begin with the decision, not the technology	148	13.6	Hypotheses need failure conditions	151	13.11	The pilot ends in a decision meeting	154
13.3	Five kinds of uncertainty	149	13.7	Baseline and comparison are operating assets	151	13.12	A stop can be a successful pilot decision	154
13.4	When model choice matters, compare system variants	149	13.8	Test the ordinary system	152	13.13	Pilots can become avoidance	155
			13.9	Inside a pilot that produces a decision	152			

THE PILOT HAS acquired a bad reputation because organizations often use it to avoid choosing.

A team demonstrates that a model can perform a task. The demonstration is declared successful. The pilot receives another extension while integration, economics, ownership, support, and operating consequences remain unresolved. Months later, the organization has learned that the technology is interesting and little else.

A pilot is not a small deployment and not an extended demonstration. It is a bounded experiment designed to reduce uncertainty for a named management decision.

Begin with the commitment management is considering. Design the pilot to produce the evidence that commitment requires.

13.1 Technical test, proof of concept, pilot, and deployment

The labels are often used loosely, but their management jobs differ.

A technical test asks whether a component performs under defined conditions. A proof of concept asks whether an idea can work well enough to justify further investment. A pilot asks whether a bounded intervention can operate inside real work and what the evidence authorizes next. Deployment places the intervention into routine operation under a current authorization.

The distinction prevents technical feasibility from being asked to settle an operating decision it was never designed to answer.

13.2 Begin with the decision, not the technology

The pilot design should start with a sentence such as: “If the intervention produces this evidence, management will authorize this next commitment.”

The next commitment may be integration work, a larger population, a new autonomy level, a procurement decision, workforce redesign, or termination. The more consequential and difficult to reverse the commitment, the stronger the comparison and evidence burden should become.

Starting from the decision also reveals when a pilot is premature. If management cannot say what it would do with a positive or negative result, the organization does not yet have an experiment. It has activity.

■ IN PRACTICE

Open every pilot charter with one sentence: if the intervention produces this evidence, management will authorize this next commitment.

DECISION

Begin with the decision the evidence will authorize. If management cannot say what it would do with a positive or negative result, it does not yet have an experiment.

13.3 Five kinds of uncertainty

A pilot may reduce capability uncertainty: whether the model performs the task with the required quality. It may reduce workflow uncertainty: whether the output fits the process and changes the constraint. It may reduce human and behavioural uncertainty: whether people use, challenge, override, or work around the system. It may reduce consequence and control uncertainty: whether errors are detected, bounded, remedied, and escalated. It may reduce economic and scaling uncertainty: whether value survives full cost, support, volume, and ordinary operating conditions.

The categories matter because one experiment cannot reduce every uncertainty equally. A technically controlled evaluation may answer whether the model can classify documents accurately. It may say little about whether employees will use the output correctly, whether review demand will overwhelm specialists, whether customer behaviour changes, or whether the economics survive at volume.

The pilot should therefore be designed around the uncertainty that blocks the next decision. When management asks one small experiment to prove capability, workflow fit, economics, control, scale, and strategic value at once, the likely result is a vague success story rather than useful evidence.

One pilot should not be forced to answer every question. The test becomes incoherent when the team changes the model, interface, workflow, population, training, staffing, and metric at the same time, then cannot explain why the result moved.

Name the primary uncertainty. Preserve the others as boundaries or secondary observations. Sequence further experiments only where the next decision requires them.

■ KEY TERM

Five uncertainties

Capability, workflow, human behaviour, consequence and control, and economics and scaling. A pilot is designed around the one blocking the next decision.

13.4 When model choice matters, compare system variants

A pilot can test the wrong thing by fixing one model configuration and asking only whether it works. Where model choice is material, compare plausible system variants: smaller versus stronger models, different routing or retrieval designs, deterministic

checks, or a non-AI path. Record the configuration rather than allowing the endpoint to drift during the experiment.

The measures should connect technical performance to the workload: accepted-output quality, difficult-case performance, latency, throughput, consumption, tool calls, retries, review and correction time, exception burden, and version stability. The end decision may legitimately be to use a smaller model, add escalation, redesign the workflow, change supplier, return to a non-AI option, hold, or stop. “Scale the model we started with” is not the default success state.

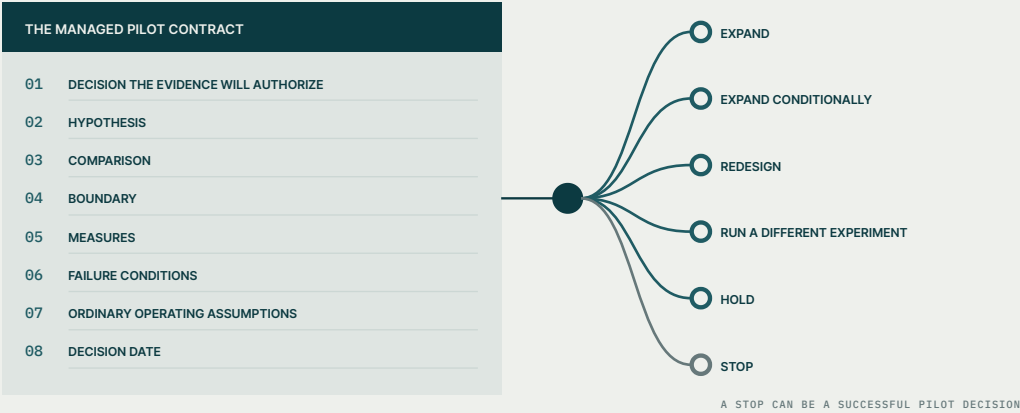
13.5 The Managed Pilot Contract

The contract is valuable because it turns enthusiasm into a falsifiable management proposition. It names the decision, hypothesis, comparison, boundary, measures, failure conditions, ordinary operating assumptions, decision date, and the actions management is willing to take when the evidence arrives.

That last point matters. A pilot that can only end in “continue learning” has no decision design.

FIGURE 13.1

The Managed Pilot Contract



A pilot reduces named uncertainty inside a bounded comparison and ends in a decision.

SOURCE: AUTHOR.

The full contract belongs in the operating toolkit. The doctrine belongs here: a pilot is a temporary authorization to learn, not a temporary escape from accountability.

13.6 Hypotheses need failure conditions

Teams naturally write hypotheses that describe success. “The assistant will reduce processing time by 20 per cent.” That is incomplete.

The pilot should also state what would make the result unusable: quality below the minimum standard, review burden above available capacity, a complaint pattern, a control that fails under ordinary conditions, a cost that erases the benefit, or a distributional effect the average conceals.

A failure condition makes it possible to stop without rewriting the purpose of the pilot after the evidence arrives. It protects the decision from sponsor attachment and sunk cost.

■ IN PRACTICE

Write the failure conditions before launch – quality floor, review ceiling, complaint pattern, cost limit – so the purpose cannot be rewritten afterward.

13.7 Baseline and comparison are operating assets

Without a credible comparison, improvement can be attributed to the intervention when it came from seasonality, staffing, demand, policy change, learning, or another programme. The comparison does not always need to be a randomized trial. It does need to be strong enough for the decision.

A team can use a matched process, alternating weeks, phased rollout, historical baseline adjusted for volume, parallel human workflow, or other design suited to the context. The essential question is what would probably have happened without the intervention.

The comparison should include burden as well as benefit: review time, rework, exceptions, support, integration, training, complaints, and displaced work. Gross task gains are not the same as net operating value.

■ EVIDENCE

Home Office pilots: about 23 minutes saved per transcript review and 37 per policy search; 9 per cent of summaries removed for inaccuracy or omission.

CONTEXT · EVIDENCE

UK Home Office: what a small pilot can establish

The UK Home Office’s asylum decision-support pilots show what a bounded evaluation can establish without being asked to prove more.¹ One tool summarized long asylum interview transcripts; another helped decision-makers find and summarize country-policy

information. Both were designed as aids. They did not replace the decision-making process, and decision-makers retained access to the underlying material.

The small-scale pilots used comparison groups and measured time, quality, user experience, confidence, and possible unintended consequences. The published evaluation estimated average savings of about 23 minutes when reviewing transcripts and 37 minutes when searching for policy information. It found no detectable positive or negative effect on decision quality in the tested sample.

The limitations belong inside the result. Technical specialists reviewed every generated summary and removed 9 per cent because of inaccuracy or missing information. Twenty-three per cent of users were not fully confident in summaries that proceeded. The appropriate conclusion was therefore not unrestricted readiness. It was that the tools showed bounded efficiency potential while limitations, monitoring, and further evaluation remained necessary before a wider rollout.

13.8 Test the ordinary system

A pilot should deliberately expose the real workflow and operating environment to representative inconvenience. Include ordinary staffing, realistic volumes, routine handoffs, common exceptions, and the periods when the sponsor is not standing beside the team. Otherwise the pilot measures the performance of an exceptional project environment rather than the AI use management intends to scale.

The failure conditions should include organizational failure as well as model failure: review queues exceed capacity, required data arrive too late, users bypass the workflow, specialists cannot resolve exceptions within the safe period, or the cost of correction consumes the expected benefit.

The pilot should not be protected by conditions that disappear at launch. Use ordinary staff, ordinary queues, expected edge cases, real handoffs, and realistic support. Include a peak period where volume matters. Rotate reviewers. Remove the developer from daily rescue. Observe what happens when the expert is unavailable.

The aim is not to make the test artificially harsh. It is to prevent exceptional attention from being mistaken for production readiness.

13.9 Inside a pilot that produces a decision

The following procurement scene is illustrative.² It is designed to show the decision logic, not to prescribe one evaluation method for every finance process.

A manufacturing group's shared-services team receives thousands of supplier invoices each month. Staff spend substantial time classifying exceptions: missing purchase order, quantity mismatch, suspected duplicate, tax issue, supplier-master problem, or simple

data-entry error. The proposed AI assistant will classify the exception, retrieve the relevant record, and draft the supplier or internal follow-up. It will not approve payment or change supplier data.

The sponsor wants to know whether to fund integration into the accounts-payable platform. The pilot therefore does not ask only whether the classifier is accurate. It asks whether the intervention reduces the age and handling cost of the exception queue without increasing incorrect routing, duplicate work, supplier complaints, or control failures.

The team defines a six-week comparison across two similar processing groups. Both groups continue using the same payment controls. One receives AI classification and drafted follow-up; the other uses the existing process. The pilot includes month-end volume, new suppliers, incomplete purchase orders, and rotating staff. A finance-control reviewer who does not report to the sponsor samples decisions independently.

The hypothesis is deliberately bounded. Standard duplicate, quantity-mismatch, and missing-document cases should be classified with enough quality to reduce first-touch time and queue age. Failure conditions include any AI-assisted action that bypasses payment control, materially higher incorrect routing, review burden that consumes the saved time, or a supplier-complaint pattern linked to poor drafts.

At the decision meeting, the headline result looks positive. First-touch time falls sharply for standard duplicates and missing documents. Staff appreciate the retrieved context. The quality reviewer finds no bypass of payment controls.

But the aggregate obscures an important weakness. Missing-purchase-order cases are routed faster to procurement, where they wait longer because ownership is unclear. Accounts payable appears more productive while the enterprise queue has moved. The pilot has improved one departmental metric without resolving the workflow constraint.

The economic result is also partial. Gross handling time falls, but reviewers still inspect nearly every drafted supplier message because the tone and requested evidence vary by region. The organization has created capacity potential, not yet realized cost savings.

The team does not call the pilot a failure. It makes a narrower decision. Integration is approved for classification and evidence retrieval in the three stable categories. Drafted external communication remains human-led. Missing-purchase-order cases are excluded until procurement ownership and service levels are redesigned. A second experiment will test whether the new handoff reduces total queue age, not merely accounts-payable handling time.

The pilot has done its job. It did not prove that “AI works in accounts payable.” It produced evidence precise enough to authorize one commitment and reject another.

13.10 Evaluation needs enough independence

Independence is particularly important when the pilot sponsor controls both the intervention and the success metric. The evaluator does not need to be hostile or external in every case. It needs access to the baseline, the right to inspect excluded cases, permission to preserve negative findings, and enough authority to say that the evidence does not support the preferred next step.

This is one reason counterevidence is valuable. A pilot that exposes a weak assumption and prevents an expensive deployment may have created more value than one that produces a flattering metric and leaves the real uncertainty unresolved.

The sponsor should not be the only person deciding what the evidence means. Sponsors carry urgency, investment, reputation, and a preferred outcome. That does not make them dishonest. It makes independent challenge necessary.

□ WATCH FOR

Pilots evaluated only by the sponsor who controls both the intervention and the success metric.

13.11 The pilot ends in a decision meeting

Pilot evidence also expires. A result tied to one model, population, volume, incentive structure, or workflow cannot be carried indefinitely into a changed system. The decision record should state what change would require a new test.

The final meeting should not become a presentation of everything the team accomplished. It should answer four questions: What did we learn? What remains uncertain? Which commitment does the evidence support? What happens to the intervention now?

The available decisions should be known before the pilot starts: expand, expand conditionally, redesign, run a different experiment, hold, or stop. Every option needs an owner and a record.

13.12 A stop can be a successful pilot decision

McDonald's decision to end its IBM automated order-taking test is a useful counterweight to the idea that a pilot creates value only when it scales.³ The company ended the tested configuration and removed it from more than 100 trial restaurants, while continuing to say that voice ordering could have a future and that other approaches would be evaluated.

The public record establishes the management distinction that matters here: declining to scale one configuration is not the same as rejecting the capability category. It does not disclose the complete internal scorecard, exact economics, or one definitive mechanism behind the stop. A pilot can create value by preventing a larger commitment, forcing a redesign, or changing suppliers. The decision is the output.

13.13 Pilots can become avoidance

A pilot becomes avoidance when the organization keeps renewing the experimental label after the uncertainty that justified it has already been resolved. Sometimes the evidence supports deployment and management is reluctant to own the consequence. Sometimes the evidence is weak and the sponsor is reluctant to stop. In both cases, “pilot” becomes a way to postpone a decision.

Experimentation can be used to delay a difficult decision. Teams request more data after the original threshold is met. Sponsors extend a weak pilot because stopping would embarrass the programme. Governance requires another round without identifying what uncertainty it will reduce. The organization accumulates pilots that are permanently promising and never operational.

The remedy is not fewer experiments. It is time-boxed experiments with explicit decisions. A pilot should close when its decision has been made, even if the organization chooses another pilot for a different uncertainty.

DECISION

A pilot is successful when it changes what management is justified in doing – not when the demonstration is impressive.

KEY TAKEAWAYS

- A pilot is a bounded experiment that reduces uncertainty for a named management decision.
- Hypotheses need failure conditions, and comparisons must include burden as well as benefit.
- A stop can be a successful pilot decision. A pilot renewed past its uncertainty is avoidance.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *What decision will this pilot's evidence authorize?*
- 02 *Who, other than the sponsor, can say the evidence does not support the next step?*
- 03 *Which of our pilots has outlived the uncertainty that justified it?*

14

CHAPTER 14 • PART IV

Keep Value and Permission Current

THE QUESTION

Do the conditions that justified approval still hold now?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Launch is time-stamped	Six signal families	Every material signal has a contract	Change control reopens decisions	Economics can reverse the case

DECISION RULE

Continue, restrict, reauthorize, and exit are all management decisions.

THE FIVE CONDITIONS • WHERE THIS CHAPTER SITS

Value clarity	Workflow fit	Operating rules	Ownership and accountability	Evidence and learning
---------------	--------------	-----------------	------------------------------	-----------------------

IN THIS CHAPTER

14.1 Launch is a time-stamped authorization	158	14.5 Every material signal needs a contract	160	14.9 Inside a runtime incident and renewed approval	162
14.2 Monitoring is necessary but not sufficient	158	14.6 Assurance runs on several clocks	161	14.10 Runtime economics can reverse the value case	163
14.3 Begin with what was authorized	159	14.7 Change is not automatically failure	161	14.11 Proportionate assurance	164
14.4 Six signal families	160	14.8 A model-economics change contract	161	14.12 Permission is a living management decision	164

IN 2021, ZILLOW PAUSED new contracts for Zillow Offers and later announced that it would wind the business down.¹ The company pointed to forecasting volatility and an inability to manage the operational scale and uncertainty of the programme as expected.

The important lesson is not that one model prediction failed. It is that a value case can reverse after launch. Market conditions move. Capacity constraints emerge. Costs change. Error patterns become visible. The system that looked viable at authorization can become uneconomic or unmanageable in operation.

Launch is therefore not the end of the management decision. It is the date on which the organization accepted a particular use, under particular evidence, boundaries, assumptions, and conditions.

Permission must remain current.

Deployment is a time-stamped authorization. The organization earns the right to continue by observing what changes and connecting signals to action.

14.1 Launch is a time-stamped authorization

That matters because every part of the system can change after launch: the model, vendor, data, population, volume, upstream process, downstream team, law, risk tolerance, economics, or organizational objective. Runtime assurance is the mechanism that keeps permission connected to those changing conditions.

14.2 Monitoring is necessary but not sufficient

Organizations often respond to runtime uncertainty by adding dashboards. A dashboard can show that something moved. It cannot decide whether the movement matters, who owns the response, what temporary state is safe, or whether the authorization still holds.

A monitored system can remain unmanaged. Signals accumulate. Alerts become noise. Review meetings note the pattern. Nobody has the authority, evidence, or response contract to change the intervention.

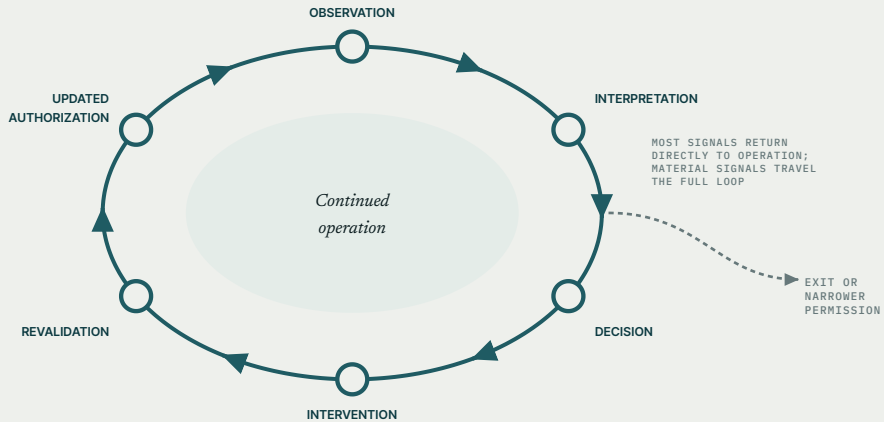
BOUNDARY

Monitoring is necessary but not sufficient. A monitored system can remain unmanaged.

Assurance begins when observation is connected to interpretation, decision, intervention, revalidation, and an updated authorization or exit.

FIGURE 14.1

The Runtime Assurance Loop



Observation matters only when signals lead to interpretation, action, revalidation, and an updated authorization or exit.

SOURCE: AUTHOR.

The loop is not a ceremonial cycle. Most observations should return directly to continued operation. Material signals require a different path because the prior decision is no longer sufficient on its own.

■ KEY TERM

Runtime assurance

Observation connected to interpretation, decision, intervention, revalidation, and an updated authorization or exit.

14.3 Begin with what was authorized

Runtime judgment requires a reference point. The authorization should identify the approved purpose, users, affected population, workflow placement, data, action rights, autonomy, quality and value thresholds, review and remedy, volume, geography, model and vendor version, fallback, material-change triggers, and expiry.

Without that record, drift becomes difficult to name. The organization knows that the system is different but not whether the difference violates a condition management previously relied upon.

The authorization record also prevents a quiet expansion of purpose. A tool approved to summarize internal notes can become a source of customer advice. A recommendation can become default action. A pilot population can become enterprise access. Each change

may appear small. Together they can create a system management never actually authorized.

14.4 Six signal families

The runtime record should cover six families of signal.

Value and economics ask whether the intended outcome still improves after current cost and burden. System integrity covers availability, data pipelines, security, vendor dependence, and component failure. Performance and quality track accuracy, reliability, calibration, difficult cases, and distributional variation. Workflow and people reveal queue movement, workarounds, review capacity, adoption, skill, and hidden labour. Consequence and control cover complaints, incidents, rights, remedies, overrides, and control effectiveness. Change and evidence age identify model updates, policy changes, data shifts, market movement, new populations, and altered assumptions.

No organization needs every possible metric. It needs the signals that can change the decision.

□ WATCH FOR

Summaries that become customer advice, recommendations that become default action, and pilot populations that become enterprise access.

14.5 Every material signal needs a contract

Some signals justify immediate containment; others justify investigation, a narrower boundary, more evidence, or no action at all. The contract prevents every anomaly from becoming a crisis and every warning from becoming background noise.

A signal without an owner, threshold, cadence, interpretation rule, and response is not assurance. It is data collection.

For each material signal, the organization should know what is measured, why it matters, who sees it, how often, what normal variation looks like, what triggers investigation, who may change the operating state, what evidence is preserved, and how the outcome updates the authorization.

■ IN PRACTICE

Give every material signal an owner, threshold, cadence, interpretation rule, safe state, and response – or stop calling it assurance.

This is especially important for delayed signals. Confirmed fraud, customer harm, model bias, revenue impact, and long-term quality may arrive after the operational dashboard has declared success. Leading and lagging evidence must remain connected.

14.6 Assurance runs on several clocks

Trying to force all of those signals into one review cadence either overwhelms the organization or allows fast-moving problems to wait for a slow meeting.

Some signals need continuous observation: service failure, unauthorized action, security event, or loss of fallback. Others belong to daily or weekly operating review: queue age, correction, complaint, review burden, and unit cost. Strategic value may need monthly or quarterly evaluation. A vendor or policy change can reopen the decision immediately regardless of the normal cadence.

The mistake is to place every issue on one committee calendar. A harmful action cannot wait for a quarterly value review. A long-term outcome should not be judged from one day of variation. The consequence determines the clock.

14.7 Change is not automatically failure

A system that needs modification has not necessarily failed. Adaptation is part of operating AI. Models, data, workflows, prices, policies, customer behaviour, and organizational priorities change.

The failure is allowing change to bypass decision control. A vendor updates the model, the interface looks the same, and the prior approval is assumed to remain valid. A business unit adds a new population without testing. A workflow removes a human checkpoint because the queue is slow. A cost increase quietly erases the value case.

Change control should classify what can proceed routinely, what requires targeted revalidation, and what reopens the full authorization. The classification should depend on whether the change can alter value, consequence, rights, evidence, or recovery.

The remedy may therefore be technical, operational, or decisional: repair the model, redesign the workflow, narrow the boundary, update the evidence requirement, or reauthorize the use under a different condition. Runtime assurance exists to make that distinction explicit.

14.8 A model-economics change contract

Some technical changes are economic changes even when the user interface looks identical. A material change in price or billing method, model or version, depreciation date, rate limit, latency, context configuration, retrieval source, tool behaviour, data route,

vendor term, deployment arrangement, or migration condition can alter the evidence that justified the intervention.

The runtime record should therefore name which changes reopen the decision automatically. The response may be a targeted revalidation rather than a full reassessment, but silence should not count as continued authorization. Where dependence on one vendor, model family, or endpoint is material, portability and fallback evidence belong in the value case because migration difficulty is part of the operating cost.

■ IN PRACTICE

Align your change classification with your vendors'. An update the vendor calls routine can be material to your authorization.

The Department for Transport's consultation-analysis evaluation provides a practical example: changes to prompts and model versions required re-engineering and re-evaluation of the pipeline rather than automatic inheritance of the prior evidence.² The principle is broader than any one provider. A configuration change that can alter quality, burden, or economics can also alter permission.

14.9 Inside a runtime incident and renewed approval

The following scene returns to the illustrative card-protection intervention from Chapter 12. It shows the lifecycle after the bank has completed further testing and authorized a narrow form of automated action.

Six months after the readiness review, the bank permits the system to place short automatic holds only when two independent high-confidence signals agree. The action is limited to consumer debit cards, defined transaction types, and a maximum hold period. Customers can confirm a legitimate purchase in the mobile app or reach an agent with authority to release the hold. The authorization excludes business cards, certain travel patterns, and transactions where identity confidence is weak.

The system performs within its expected range until a vendor update changes how transaction velocity is calculated for mobile-wallet tokens. During a holiday travel period, complaint volume rises and the call queue lengthens. The model dashboard initially looks normal because confirmed-fraud labels arrive later. The first decisive signal comes from operations: legitimate customers are taking longer to regain access to funds, and a growing share of cases involves the same mobile-wallet pattern.

The signal contract triggers investigation and gives fraud operations authority to move the system to a safe state. Autonomous holds are suspended for mobile-wallet transactions. The system continues ranking alerts, and analysts review them manually. Customer service receives a priority release queue and a script that acknowledges the issue without attributing blame before the investigation is complete.

The team preserves the affected transactions, model version, feature values, release times, complaints, and overrides. Engineering identifies the changed velocity feature. Fraud analysts find that the update improved detection in one segment but created a new false-positive pattern for legitimate travellers. The review also exposes a governance gap: the vendor had classified the update as routine, while the bank's authorization treated changes to decision logic as material. The notification contract and internal change trigger were misaligned.

The remedy is not merely to lower a threshold. The bank restores the prior feature logic, adds segment-level monitoring for mobile-wallet and travel patterns, changes the vendor-change classification, and tests the release process under peak volume. A shadow run compares the revised system with analyst decisions. Customer remedy confirms that release times have returned below the required service level and that no segment remains outside tolerance.

Revalidation supports narrower renewed approval. Automatic holds resume for the original transaction types, but mobile-wallet cases remain recommendation-only until a longer observation period clears. The updated record names the new monitoring signal, vendor notification requirement, safe-state authority, and expiry date.

Nothing in the scene depends on pretending that incidents can be eliminated. The management achievement is that the organization detected a consequence the model dashboard lagged, changed the operating state quickly, found the mechanism, repaired both technology and governance, tested the remedy, and altered permission rather than simply declaring the system fixed.

I would rather see an intervention restricted for a week than discover months later that everyone watched the dashboard while the customer absorbed the error.

14.10 Runtime economics can reverse the value case

A system can remain technically healthy while becoming economically irrational. Demand may fall. The vendor may change pricing. Review burden may grow as easy cases are automated and difficult cases concentrate. Integration support may become permanent. A supposedly avoided headcount cost may never materialize because the released capacity is absorbed elsewhere.

Runtime assurance therefore has to preserve the economic mechanism, not merely model performance. Permission should be able to narrow or end when the business case changes even if the AI continues to function exactly as designed.

Value assurance should therefore revisit the original comparison. Is the intervention still better than the realistic alternative? Are benefits realized or merely potential? Which

costs and adverse consequences have appeared since launch? Does the strategic option still matter?

Stopping a use whose value case no longer holds is not evidence that the original decision was foolish. It can be evidence that management remained responsive to reality.

■ IN PRACTICE

At every renewal, re-price the realistic alternative. Continuation must beat it after current cost.

14.11 Proportionate assurance

Too much assurance can damage the system it protects. Excessive sampling, duplicate review, and repeated approval can erase the benefit, overload specialists, and drive users toward ungoverned alternatives.

The answer is proportionality. High-consequence actions, weak reversibility, uncertain performance, changing populations, and concentrated harm require stronger observation and faster response. Stable, low-consequence uses operating inside approved components can follow lighter renewal and event-triggered review.

14.12 Permission is a living management decision

The organization should always be able to answer: What is running now? Under which authorization? What changed since approval? Which signals can narrow or stop it? Who holds that authority? What evidence supports continuation? When does the decision expire?

When those answers disappear, deployment becomes inertia. The system continues because it exists, not because management has concluded that it should.

AI Value Management treats operation differently. Every live intervention remains connected to its value case, workflow, authority, evidence, and current conditions. Continue is a decision. Restrict is a decision. Reauthorize is a decision. Exit is a decision.

That discipline also changes the portfolio. Capital and specialist attention should not remain trapped in interventions whose evidence has weakened while stronger opportunities wait. Part V takes the decision system from one intervention to the enterprise portfolio and the organization capable of repeating it.

DECISION

Do not ask whether the system was approved. Ask whether the evidence and conditions that justified approval still hold now.

KEY TAKEAWAYS

- Launch is a time-stamped authorization. Permission must remain current.
- Every material signal needs an owner, threshold, cadence, and response.
- Change control and runtime economics can reopen, narrow, or end the decision.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *What is running now, under which authorization, and what has changed since approval?*
- 02 *When a vendor classifies an update as routine, does our authorization agree?*
- 03 *Is each live intervention still better than its realistic alternative after current cost?*



PART V

Repeating the Discipline

Parts II to IV followed the management of specific AI-enabled interventions from value case and workflow design through readiness, piloting, and runtime authorization. Part V asks whether the enterprise can repeat that discipline when many interventions compete for the same money, experts, platforms, attention, and tolerance for consequence.

This is where isolated good practice either becomes an organizational capability or disappears into project history. The test is not whether one team can manage AI well. It is whether the enterprise can choose what deserves commitment, stop what no longer does, and make better decisions easier to repeat.

A managed portfolio changes commitments. A capable organization makes those changes executable without removing accountability from the business.

15	Managing the AI Portfolio	168
16	Building the AI Value Organization	178

15

CHAPTER 15 • PART V

Managing the AI Portfolio

THE QUESTION

Which commitment deserves the next scarce unit of capacity?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
An inventory is not allocation	Learning capacity is the constraint	Three classes, three logics	Fund the next credible commitment	Stopping releases value

DECISION RULE

A portfolio is managed only when the review changes a commitment.

THE FIVE CONDITIONS • WHERE THIS CHAPTER SITS

Value clarity	Workflow fit	Operating rules	Ownership and accountability	Evidence and learning
---------------	--------------	-----------------	------------------------------	-----------------------

IN THIS CHAPTER

15.1	An inventory cannot make the trade-off	169	15.5	Public claims become harder to interpret at portfolio level	172
15.2	The scarce resource is often the ability to learn	169	15.6	Inside a portfolio trade-off meeting	172
15.3	Three classes, three decision logics	170	15.7	Portfolio value is not the sum of business cases	174
15.4	Evidence should govern the distance to the next funding decision	171	15.8	The stop decision releases value	175
			15.9	The portfolio review is a decision meeting	176

THE INVENTORY GREW from ten AI use cases to 126. That looked like momentum. It did not prove that the portfolio was being managed.

A 2026 review by the United States Government Accountability Office found that the Internal Revenue Service had expanded its recorded AI use cases substantially.¹ The same review found missing benefit explanations, incomplete inventory information, skills gaps, and no process that connected individual investments to agency-wide priorities. The lesson travels beyond the public sector: visibility is necessary, but a list does not allocate anything.

The scarce commitment may be money. It may also be the next engineering slot, the attention of a domain expert, access to production data, a legal review, a model-validation window, a scarce reviewer, executive sponsorship, or the patience of employees asked to redesign their work.

A portfolio is not a catalogue of AI activity. It is the enterprise decision about which interventions should expand, fund the next evidence, repair, consolidate, hold, restrict, or retire.

15.1 An inventory cannot make the trade-off

Management cannot allocate what it cannot see. The organization needs a current record of its models, vendors, agents, experiments, workflows, data dependencies, owners, operating permissions, costs, and evidence. But visibility is only the beginning.

A list of one hundred initiatives can contain live systems producing measurable value, experiments reducing strategic uncertainty, duplicate tools, abandoned pilots, required controls, shared infrastructure, and projects whose sponsors have left. Sorting them by business unit or technology does not tell management which commitment is most important or which one should be stopped.

Nor does a single score solve the problem. A live assistant with measured operating value, a speculative autonomous agent, a shared identity service, and a mandatory safety control have different purposes and evidence. Ranking all four on one blended number makes unlike decisions appear comparable.

The portfolio must first classify the commitment, then apply the decision logic appropriate to it.

15.2 The scarce resource is often the ability to learn

The allocation question is therefore comparative: which commitment deserves the next scarce unit of capacity, and what other commitment will wait because of that choice? A portfolio meeting that approves everything has avoided the portfolio decision.

BOUNDARY

A portfolio meeting that approves everything has avoided the portfolio decision.

Budget is visible. The capacities that decide whether an intervention can become valuable are often harder to see.

Domain experts must identify representative cases and judge whether outputs make sense. Operations leaders must redesign handoffs and absorb changed work. Engineers must integrate systems and preserve fallback. Finance must test the economics. Privacy, security, legal, risk, procurement, and assurance teams have limited review capacity. Frontline employees and customers supply evidence that no laboratory can manufacture.

An intervention can clear its own financial hurdle and still be the wrong portfolio choice because it consumes the capability required by a more important commitment.

A company may be able to fund ten pilots and have enough operating attention to learn from only three. Funding all ten does not create more optionality when none receives enough evidence, integration, or ownership to reach a decision.

■ IN PRACTICE

Count learning capacity before counting budget: domain experts, reviewers, validation windows, and integration slots.

The portfolio question is therefore not merely whether an initiative has a positive expected return. It is whether this is the best use of the resources and dependencies the organization cannot expand quickly.

15.3 Three classes, three decision logics

The AI Portfolio Allocation Map separates commitments before management allocates across them.

A current-value intervention should show realized outcomes, current economics, operating resilience, and the effects of additional scale. A strategic experiment should show the importance of the uncertainty, the cost of the next test, and the decision that test will change. A shared foundation should show reuse, reliability, avoided duplication, and the constraints it removes.

■ KEY TERM

Portfolio classes

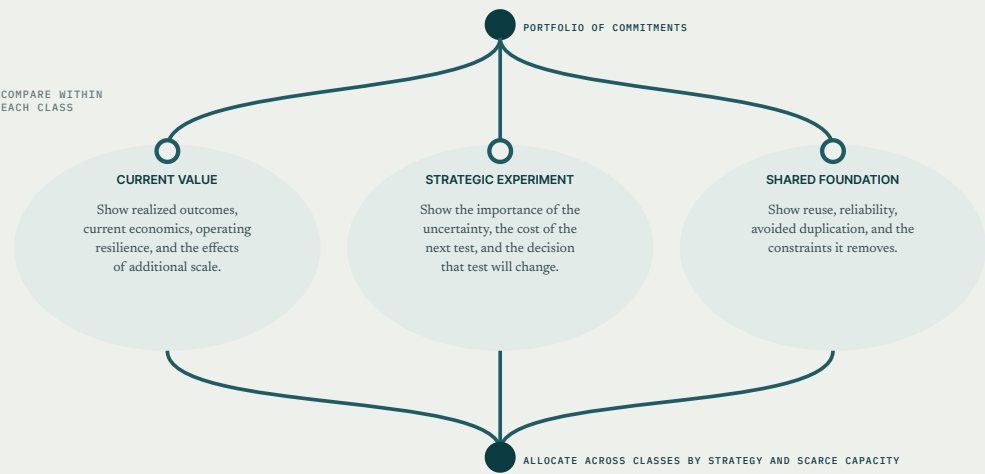
Current value, strategic experiment, and shared foundation. Compare within a class; allocate across classes by strategy and scarce capacity.

Management can then decide how much capacity should protect current value, how much should test future options, and how much should build reusable capability. There is no universal ratio. The discipline is to make the mix deliberate rather than allowing sponsorship and queue position to decide it silently.

15.4 Evidence should govern the distance to the next funding decision

FIGURE 15.1

The AI Portfolio Allocation Map



Compare commitments within their class, then allocate across classes by strategy and scarce capacity.

SOURCE: AUTHOR.

AI work often enters annual budgeting before its most important uncertainties have been reduced. The result is a large commitment justified by a distant destination rather than the evidence required for the next decision.

The portfolio should fund the next credible commitment. A strategic experiment may need six weeks to test whether a workflow can produce decision-grade evidence. A proven intervention may need one region of expansion to establish whether reviewer capacity and unit economics hold. A shared foundation may need a service level tied to named users rather than an open-ended platform ambition.

The funding interval should match uncertainty. Stable operating value can justify a longer commitment. A speculative agent whose value depends on unresolved authority, data, and workflow assumptions should return sooner.

Capacity must be allocated at the same time as money. Approving a pilot without domain review, data support, assurance, or a production path

■ IN PRACTICE

Approve money and capacity together: domain review, data support, assurance, and a production path.

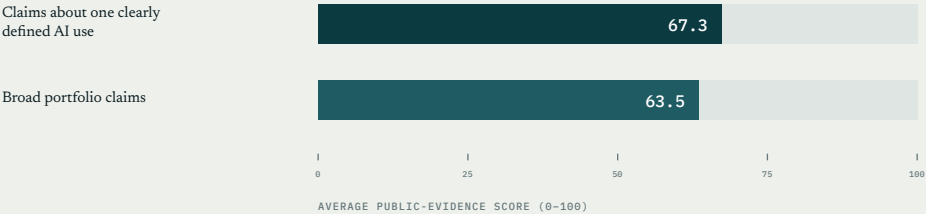
merely creates a queue whose cost is hidden from the business case.

15.5 Public claims become harder to interpret at portfolio level

Portfolio-level claims are harder to verify. In the 2026 Scalebotics benchmark, claims about one clearly defined AI use averaged 67.3 out of 100 on the public-evidence score against 63.5 for broad portfolio claims – a modest gap with a clear cause.²

EXHIBIT 15.1

Boundaries improve inspectability



Claims about one clearly defined AI use scored higher on public evidence than broad portfolio claims.

SOURCE: SCALEBOTICS RESEARCH, 2026 AI VALUE PERFORMANCE BENCHMARK.

A workflow-level claim can usually name a user, intervention, comparison, operating period, and consequence. A portfolio claim may combine use-case counts, employee adoption, expected revenue exposure, infrastructure investment, risk reduction, and annualized benefit across different boundaries. There may be no single counterfactual that makes the total interpretable.

Internal portfolio management faces the same problem. Management should not sum heterogeneous business cases and call the result enterprise AI value. It should preserve the records for each AI use, understand dependencies and shared costs, then allocate across classes according to strategy and scarce capacity.

15.6 Inside a portfolio trade-off meeting

The following scene is illustrative. It combines several kinds of AI commitment to show what a portfolio decision must do; it does not describe a named organization.³

A diversified group with banking and industrial businesses enters its quarterly portfolio meeting with twenty-seven active interventions. The meeting does not review all twenty-seven. It focuses on six decisions that compete for two production-integration slots, one model-validation team, and a limited group of legal and domain specialists.

The first request is to expand a customer-service drafting assistant from one product line to three. It has six months of bounded live evidence: handling time has fallen, quality remains above the agreed threshold, complaint volume is stable, and the review burden is declining as the workflow improves. The business owner asks for both production slots so that expansion can occur before the next peak season.

The second request comes from the procurement team. Its contract-review pilot identified missing clauses and shortened first-pass review, but performance deteriorated on non-standard agreements and the legal team found that the clause library itself contained inconsistent guidance. The pilot produced useful evidence. It did not support deployment.

The third item is a card-protection intervention like the one followed in Chapters 12 and 14. It is producing measurable fraud benefits, but the mobile-wallet incident revealed a vendor-change and segment-monitoring weakness. The live authorization remains narrower than the business wants. Remediation needs one validation slot and engineering work before the boundary can grow.

The fourth is a predictive-maintenance proposal from an industrial site that – unlike the plant in Chapters 8 and 10 – has not yet established a reliable failure baseline. Its expected value is strategically attractive, but the team has not yet secured enough maintenance-engineer time to label representative cases. The request is for a broad data-platform build and a full pilot.

The fifth is a shared evaluation and observability service used by nine interventions. Its team needs capacity to support version comparison, incident reconstruction, and common evidence records. The benefit does not sit neatly inside one workflow, but each new intervention is rebuilding part of the same capability because the shared service is underfunded.

The sixth is a second customer-chatbot platform sponsored by another business unit. Its local forecast is positive, but it duplicates procurement, identity, training, logging, and monitoring already provided elsewhere. Its main advantage is a different vendor relationship, not a distinct customer capability.

The meeting begins badly. Each sponsor presents the intervention as though the other five do not exist. Every business case is positive. Every deadline is urgent. The portfolio lead changes the question. Which commitments must the organization protect, which uncertainties are worth reducing now, and which dependency is constraining several decisions at once?

The group protects the card-protection remediation because current value and customer consequence are both material. It receives the model-validation slot and one integration slot. The drafting assistant receives a narrower expansion to one additional

product line, using the remaining integration slot, with a stated capacity threshold before the third product can enter.

The procurement system is not labelled a failed pilot. Its next commitment changes. Funding goes first to cleaning and governing the clause library and constructing a representative non-standard test set. No production integration is authorized.

The maintenance proposal receives a small evidence budget rather than a platform programme. The business must establish the baseline, identify the constraint in the maintenance workflow, and secure named domain capacity before returning. The duplicate chatbot is stopped; the sponsor may propose a bounded resilience test if vendor diversity is the real strategic objective.

The shared evaluation service receives capacity because it removes a constraint across the portfolio. Its service level, users, cost, and evidence are made explicit so that “shared platform” does not become an exemption from accountability.

The value of the meeting is not that every sponsor leaves happy. It is that the organization leaves with a coherent set of commitments, explicit non-commitments, and released capacity. A portfolio with no rejection, delay, consolidation, or exit is not allocating. It is accumulating.

15.7 Portfolio value is not the sum of business cases

Individual value cases interact. Two assistants may claim the same employee time saving. A shared platform cost may be counted inside every intervention or in none. One automation may remove work that another project was expected to improve. Several systems may depend on the same vendor, retrieval index, identity service, reviewer group, or data source.

Adding every forecast benefit can double count value. Adding every project cost can double count infrastructure. Treating each use as independent can hide concentration and a single point of failure.

The portfolio must separate intervention-specific and shared value, allocate common cost transparently, record transferred workload, identify whether benefits are additive or overlapping, and test whether the combined system fits available operating capacity.

Concentration may be accepted deliberately to gain speed, consistency, or purchasing power. It should not arise accidentally because teams selected the same vendor without seeing the enterprise dependency.

Model, vendor, and deployment concentration should therefore be visible at portfolio level. Shared routing and evaluation services, common endpoints, rate limits, and migration capacity can become enterprise constraints even when each intervention has a

plausible local value case. A shared model service may reduce duplicated cost and improve control; it may also concentrate exposure.

EXHIBIT 15.2

Portfolio Trade-off Record

COMMITMENT	CLASS	DECISION
Card-protection remediation	Current value	Protect: model-validation slot and one integration slot
Customer-service drafting assistant	Current value	Narrower expansion to one additional product line, with a capacity threshold
Contract-review pilot	Strategic experiment	Fund clause-library cleanup and a non-standard test set; no production integration
Predictive-maintenance proposal	Strategic experiment	Small evidence budget; return with baseline, constraint, and named domain capacity
Shared evaluation and observability service	Shared foundation	Fund capacity with an explicit service level, users, cost, and evidence
Second customer-chatbot platform	Duplicate	Stop

A portfolio meeting changes commitments rather than merely reporting activity.

SOURCE: AUTHOR.

The portfolio decision is to make that trade-off explicit. Procurement should not become invisible technical infrastructure that commits many workflows to one vendor, version path, data route, or migration burden without the same allocation attention given to other shared foundations.

WATCH FOR

Many teams selecting the same vendor, endpoint, or data route without anyone seeing the enterprise dependency.

15.8 The stop decision releases value

Stopping is an allocation action. It releases money, specialist capacity, data access, integration slots, management attention, and sometimes customer or workforce burden. A portfolio that cannot stop weak commitments will eventually crowd out better ones, even if every individual project can tell a plausible story about why one more quarter might help.

Organizations celebrate launches and often treat retirement as embarrassment. That creates portfolio drag.

A weak intervention can remain alive through a small licence, an integration, a review queue, and a sponsor who still believes the next model release will rescue it. The invoice

may be modest. The hidden cost includes security, procurement, monitoring, data maintenance, training, incident response, and attention that could support a stronger commitment.

A well-managed exit preserves records, customer and employee continuity, data obligations, learning, and dependencies, then returns scarce capability to the portfolio.

■ IN PRACTICE

Run retirement as a managed release: preserve records, obligations, and learning, then return the capacity to the portfolio.

A stopped experiment can be valuable when it prevents a larger commitment based on a weak premise. A retired live system can be evidence of good management when the value case, conditions, or alternatives changed.

15.9 The portfolio review is a decision meeting

The portfolio record earns its place only when it changes a commitment. Reporting the number of use cases, aggregate spend, or percentage of employees using AI may describe activity; it does not allocate scarce capacity. The meeting should leave with something expanded, narrowed, delayed, consolidated, funded for evidence, or stopped.

Status is useful only when it prepares an action. For each material commitment, management should know its class, present boundary, value or uncertainty, current evidence, full cost, constrained capacity, dependencies, changes since the previous decision, and the action required now.

The action should be a verb: expand, fund the next evidence, repair, consolidate, hold, restrict, or retire.

The record should preserve dissent. Finance may question the value estimate. Operations may believe the review burden is understated. Risk may identify concentration. The business owner may argue that a strategic window justifies a bounded exception. A green status should not erase those differences.

The portfolio must also be rebalanced when evidence, strategy, cost, vendors, incidents, or operating conditions change. An annual label is not enough for an intervention whose uncertainty will resolve in six weeks or whose authorization can be altered by a model update tomorrow.

DECISION

A portfolio is managed only when the review changes a commitment: expand, fund the next evidence, repair, consolidate, hold, restrict, or retire.

Once those choices are visible, the final organizational question becomes unavoidable. Who keeps the system coherent, and who makes the decisions executable without allowing the centre to absorb accountability from the business?

KEY TAKEAWAYS

- An inventory provides visibility. It does not allocate anything.
- Classify commitments – current value, strategic experiment, shared foundation – before allocating across them.
- Stopping is an allocation action. It releases money, specialists, and attention.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *How many interventions can we actually learn from this quarter?*
- 02 *Which shared dependency is constraining several decisions at once?*
- 03 *What did our last portfolio review expand, narrow, consolidate, or stop?*

16

CHAPTER 16 • PART V

Building the AI Value Organization

THE QUESTION

Who keeps the system coherent without absorbing accountability?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Five kinds of repeatable work	A minimum viable operating model	The office owns the system, not the result	The operating contract creates speed	Incentives reveal the real organization

DECISION RULE

Make accountable AI decisions repeatable, and revise them when conditions change.

THE FIVE CONDITIONS • WHERE THIS CHAPTER SITS

Value clarity	Workflow fit	Operating rules	Ownership and accountability	Evidence and learning
---------------	--------------	-----------------	------------------------------	-----------------------

IN THIS CHAPTER

16.1 The organization exists to make five kinds of work repeatable	179	16.6 Capability development belongs inside the workflow	183
16.2 A minimum viable operating model	180	16.7 The first ninety days	183
16.3 What the AI Value Office actually owns	180	16.8 Do not make AI exceptional forever	184
16.4 The operating contract creates speed	181	16.9 The discipline is the organization	185
16.5 Capacity and incentives reveal the real organization	182		

THE ORGANIZATION IS not the AI office.

When AI becomes strategically important, leaders often ask who should own it. The question is understandable and incomplete. No single function can own the business outcome, workflow, model, data, infrastructure, employee role, customer promise, financial case, legal obligation, and independent challenge at the same time.

Assign everything to technology and a management problem becomes a systems programme. Assign everything to risk and innovation becomes a control queue. Assign everything to a central AI executive and business leaders can send responsibility upward. Assign everything to local teams and the enterprise buys the same learning repeatedly while concentration remains invisible.

The answer is not one owner for AI. It is a coherent allocation of ownership.

DECISION

The answer is not one owner for AI. It is a coherent allocation of ownership.

The AI Value Organization is a federated management capability. Business owners remain accountable for outcomes; the enterprise supplies direction, reusable capability, independent challenge, and a system that connects evidence to decisions.

16.1 The organization exists to make five kinds of work repeatable

The exact structure will vary, but the responsibilities do not disappear.

First, enterprise leadership sets direction and allocates commitments. It decides where AI matters to strategy, which consequences the organization will not accept, how scarce capability should be distributed, and which trade-offs exceed local authority.

Second, local business owners preserve outcome accountability. The owner is not merely a sponsor who secures funding. They must be able to change the workflow, allocate operating capacity, appoint competent reviewers, accept trade-offs, remedy harm, and recommend restriction or retirement.

Third, shared enablement makes good recurring decisions reusable. Identity, authoritative data, evaluation, monitoring, procurement terms, training, safe-state mechanisms, and common records should not be rebuilt from zero by every team. Reuse must remain challengeable; a shared service that does not fit real work is simply centralized friction.

Fourth, independent functions test claims and preserve mandatory boundaries. Their role is not to become the outcome owner. It is to challenge evidence, consequence,

permissions, and whether the system operates as represented. Independence without capacity is ceremonial.

Fifth, operating evidence returns to the organization. Exceptions, incidents, successful controls, user behaviour, customer remedies, vendor changes, and economic results should alter standards, training, platforms, portfolio choices, and future workflow design.

These are the jobs that must exist somewhere. The organizational design decides where each sits and how the decisions connect.

16.2 A minimum viable operating model

Minimum viable does not mean minimal seriousness. It means the organization can perform the essential decisions without building a large central bureaucracy first. In a smaller firm, the business owner, technology lead, finance partner, and risk/privacy specialist may cover several roles. What cannot disappear is the separation of accountability, challenge, evidence, and authority needed for the decision.

The design should grow from recurring work. When the same privacy interpretation, evaluation method, vendor control, review capability, or portfolio decision appears repeatedly, it becomes a candidate for shared infrastructure. Organization follows the decision system rather than preceding it.

■ IN PRACTICE

Build a shared service only when the same interpretation, method, or control recurs. Organization follows the decision system.

Most organizations do not need a large AI bureaucracy. They need enough structure to prevent consequential decisions from disappearing between functions.

A small company may assign several responsibilities to the same people. A large enterprise may distribute them across permanent functions and embedded teams. The separation is about decision rights, not headcount.

The business owns the result. Shared functions own capabilities and decisions within their competence. Executive leadership owns the operating model and enterprise trade-offs. Independent assurance tests whether the system is functioning as claimed. A compact coordinating function – this book calls it the AI Value Office – keeps the chain visible.¹

16.3 What the AI Value Office actually owns

The name is secondary to the job. The function may sit with operations, strategy, transformation, technology, finance, a Chief AI Officer, or a shared executive office. In a smaller organization, it may be one leader and a working group. In a large enterprise, it may coordinate embedded leads across business units.

It owns the management system, not every AI result.

FIGURE 16.1

The AI Value Organization



Enterprise direction, local outcome ownership, shared enablement, and independent challenge operate as one federated system.

SOURCE: AUTHOR.

Its responsibilities include maintaining the intervention and portfolio record; running material allocation and decision forums; preserving common definitions and evidence rules; coordinating reusable methods and scarce specialist capacity; identifying duplication, concentration, missing owners, and stalled decisions; and returning operating learning to the standards and services every team uses.

It should not write every business case, approve every low-consequence use, operate every workflow, own every incident, or become the permanent delivery team.

A clean accountability test is this: when an intervention fails to produce value, can the business owner be identified? When the management system fails to identify, challenge, or change that intervention, can the coordinating function be identified? Those are different failures and should have different owners.

■ KEY TERM

AI Value Office

The compact coordinating function that keeps the chain visible. It owns the management system; business owners own the results.

16.4 The operating contract creates speed

Federation works only when the enterprise and the business know what each may decide without asking again.

A local team should be able to proceed inside a delegated boundary that specifies the workflow, population, data, vendor or model, autonomy, consequence, spend, and minimum evidence requirement. Inside that boundary, the business owner can run an approved experiment, adjust a reversible workflow, correct an operating issue, or retire a weak local use without waiting for an enterprise committee.

The enterprise reserves decisions that create shared dependency or exceed local consequence: strategic vendors, common platforms, cross-business data, regulated or high-impact uses, material autonomy, enterprise reputation, and capacity conflicts that one unit cannot resolve.

Specialist functions retain rights inside their competence. Security can block an unsafe access pattern. Privacy can require a lawful data boundary. Finance can reject an unsupported value claim. A professional or safety authority can restrict a use that violates practice standards. Those rights should be explicit before conflict arrives.

The contract is two-way. Local teams receive speed and reusable capability in exchange for maintaining current records and preserving outcome ownership. The enterprise receives visibility and learning in exchange for timely services and decisions. Delegation without evidence becomes fragmentation. Central standards without service or capacity become avoidance incentives.

16.5 Capacity and incentives reveal the real organization

An operating model has finite capacity just like the portfolio. A policy that requires specialist review does not create specialists. A promise of human oversight does not create time, competence, or authority. If every intervention reaches the same legal, privacy, security, model-validation, or domain team, assurance becomes either a bottleneck or a superficial ritual.

The portfolio should therefore carry a capacity view: which roles are constrained, which decisions are waiting, where review burden is accumulating, and which shared services are approaching failure. Management can then delay a commitment, expand a capability, reuse an existing service, redesign the evidence requirement, or release capacity by stopping weaker work.

Incentives are equally structural. Organizations often ask employees to redesign work while rewarding them for preserving current output. Teams are funded for launch, not learning. A stop decision damages reputation even when it prevents a bad deployment. Reviewers are punished for slowing throughput. Business owners receive credit for adoption while costs and consequences appear elsewhere.

The incentive system can overrule the governance system. A serious organization rewards measurable outcome improvement, useful evidence, responsible challenge, reuse,

successful redesign, and the retirement of weak commitments. It does not reward the number of agents, licences, prompts, pilots, or announcements as though activity were value.

16.6 Capability development belongs inside the workflow

Training people to use a chatbot is not a workforce strategy. Roles change when AI alters what information people see, which tasks remain, where exceptions concentrate, how quality is judged, and which decisions require human authority.

Employees need practice with representative cases, system limits, review standards, escalation, and communication with affected people. Managers need to learn how to allocate AI-assisted work, protect challenge, redesign targets, and recognize when the apparent productivity gain has moved burden elsewhere. Specialists need methods for evaluation, monitoring, incident response, and reviewing approvals after material change.

The organization must also know which expertise it needs to build, buy, borrow, embed, or retain internally. General AI fluency cannot replace domain judgment, professional accountability, or the ability to recognize when a plausible answer was produced from the wrong frame.

□ WATCH FOR

Rewards for licences, pilots, prompts, and announcements. Reward outcome improvement, useful evidence, challenge, reuse, and well-timed retirement.

16.7 The first ninety days

The first ninety days are intentionally narrow.² The organization does not need to complete a full record for every AI use it can find. It needs enough visibility to identify material commitments and then run two representative decisions end to end –

one live intervention and one proposed experiment. Those decisions reveal which shared services, expertise, records, and forums are actually missing.

That is a more useful starting point than announcing an enterprise transformation programme. The organization earns the operating model by making real decisions and then standardizing what repeatedly helps.

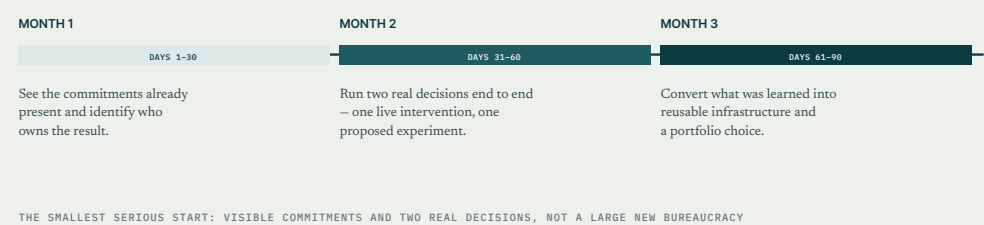
The following sequence is deliberately narrow: it is the smallest serious start for a leadership team that wants the discipline to exist in operation rather than in a slide deck.

■ IN PRACTICE

Start with two decisions run end to end – one live intervention, one proposed experiment. They show which services and forums are missing.

EXHIBIT 16.1

The First Ninety Days



The organization begins with visible commitments and two real decisions, not a large new bureaucracy.

SOURCE: AUTHOR.

The first month is not for launching more pilots. It is for seeing the commitments already present and identifying who owns the result. The second month tests whether the organization can make a disciplined decision when evidence and interests conflict. The third month converts what was learned into reusable infrastructure and a portfolio choice.

The sequence should reveal its own limits. Perhaps the organization lacks a person able to frame the problem independently of the preferred tool. Perhaps finance cannot establish a baseline. Perhaps local owners lack authority to change the workflow. Perhaps the central team can issue standards but not deliver a decision in time. These are not reasons to hide the method. They are the conditions the method was built to expose.

16.8 Do not make AI exceptional forever

Central AI teams are often created for good reasons. They aggregate scarce expertise, negotiate vendors, establish standards, and help business units begin. They fail when temporary concentration of capability becomes permanent concentration of ownership.

The central team starts collecting every use case, building every prototype, approving every deployment, operating every platform, training every employee, and reporting every benefit. Business units become customers of AI rather than owners of changed work. The queue grows and local capability remains weak.

The opposite failure is a centre that publishes guidance without authority, delivery capacity, portfolio evidence, or access to executive decisions. Business units proceed independently while the centre produces documents nobody must use.

The stronger model is proportional federation: central capability establishes the system and reusable services; local owners control outcomes and operation; embedded

specialists connect the two; independent challenge can say no, not yet, or only inside this boundary.

The coordinating office should have an expiry question. As AI becomes ordinary management, some central work should shrink or merge into strategy, operations, technology, finance, workforce, and risk. The function should not protect its existence by keeping AI exceptional forever.

■ KEY TERM

Proportional federation

Central capability builds the system; local owners control outcomes; embedded specialists connect the two; independent challenge can say no, not yet, or only inside this boundary.

16.9 The discipline is the organization

AI Value Management is not complete when leaders understand the frameworks. It becomes real when ordinary decisions embody them.

A manager defines the outcome before selecting the tool. A business owner accepts accountability for the workflow and its consequences. A reviewer has the evidence, time, competence, and authority to challenge the result. A pilot ends with a decision. A live system can be narrowed or stopped when conditions change. A portfolio meeting releases resources from a weaker commitment. A repeated exception improves the shared rule for the next team.

That is the future this book is inviting leaders to build: not an organization that merely has more AI, but one that can direct capability toward chosen outcomes without surrendering judgment, accountability, or the ability to change course.

I do not think the decisive advantage will belong to the organization that predicts every new model capability. The technology will keep moving faster than any static plan. The advantage will belong to the organization that can repeatedly decide where capability belongs, what the evidence supports, and what must change when reality disagrees.

DECISION

The AI Value Organization is not an office on the chart. It is the enterprise ability to make accountable AI decisions repeatedly and revise them when conditions change.

KEY TAKEAWAYS

- The answer is not one owner for AI. It is a coherent allocation of ownership.
- The AI Value Office owns the management system, not every result.
- The operating contract and the incentive system decide whether federation creates speed.

QUESTIONS FOR THE LEADERSHIP TEAM

- 01 *When an intervention fails, can we name the business owner – and when the system fails to catch it, the coordinating function?*
- 02 *Which decisions can local teams make without asking again?*
- 03 *Do our incentives reward outcomes and well-timed stops, or activity?*

Close

CONCLUSION

Close the Gap

THE QUESTION

What must management build so that capability becomes value?

THE ARGUMENT IN FIVE MOVES

01	02	03	04	05
Management is upstream, inside, and downstream	The gap closes one decision at a time	Advantage is decision quality under change	Begin with one real commitment	Think above the output

DECISION RULE

AI capability can be purchased. The management capacity to turn it into value must be built.

IN THIS CHAPTER

Management is not waiting at the end of the model	188	Begin with one real commitment	189
The gap closes one decision at a time	188	The organization must be able to think above the output	190
The next advantage is decision quality under change	189	Close the gap	191

THE MOST DANGEROUS AI failure is not always the one that looks broken.

A system can produce fluent, plausible, professionally acceptable work while answering the wrong question, relying on the wrong proxy, accepting an inherited assumption, or moving consequence into a part of the organization nobody is measuring. The output may survive precisely because it looks complete.

That is why this book began with a gap between capability and value. The gap exists because useful capability does not carry its own purpose, workflow, authority, economics, standard of judgment, or accountability into the organization.

Those must be built.

Management is not waiting at the end of the model

The prevailing story often places management after the technology: the model improves, the tool arrives, employees adopt it, and leaders measure what happened. That is too passive.

Management is upstream when it decides which problem is worth solving and who gets to frame it. It is inside the workflow when it determines what AI may prepare, recommend, decide, or act upon. It is present in the evidence when it defines what good looks like, which comparison matters, and what the current result is allowed to prove. It remains downstream when the organization must remedy harm, reallocate capacity, change permission, or stop the system.

AI does not enter a neutral organization. It enters objectives, incentives, data, authority, routines, and histories that management has chosen, inherited, or failed to examine. The model creates capability. The organization determines its consequence.

The gap closes one decision at a time

The unit of progress is not the number of AI initiatives. It is the quality of the next management decision around a specific use inside real work.

Where does AI belong? What outcome should change? Compared with what? Who owns the result? What may the system do? Who can challenge it? Which evidence is independent of the output itself? What is the full cost? What happens when the ordinary path fails? Which commitment does the current evidence justify?

These questions are not administrative drag around innovation. They are the means by which innovation becomes useful, answerable, and durable.

The organization does not need final proof before it learns. It needs enough evidence for the next reversible step and stronger evidence before a choice creates durable structure, customer obligation, financial exposure, safety risk, or dependence.

The next advantage is decision quality under change

This changes what management competence looks like. Leaders do not need to predict every model release. They need to build rules for hard-to-reverse commitments and enough freedom for reversible learning inside those rules. They need to know which decisions deserve friction and which should be accelerated into reality because the cost of learning is lower than the cost of continued internal debate.

The organization becomes stronger not because it reaches certainty, but because it can classify uncertainty, preserve responsibility, and change the commitment when the evidence changes.

Access to capable AI will continue to spread.¹ Features that are exceptional today will become standard. Model performance will improve, costs will change, and new forms of autonomy will enter software that organizations already use.

Access alone will therefore become a weaker source of distinction. My argument is that durable advantage will move into the organization: the ability to select better problems, redesign work, allocate authority, recognize what good looks like, generate credible evidence, adapt faster, and withdraw permission when the old conditions no longer hold.

That is not a promise that every AI investment will succeed. A mature system will stop some ideas earlier, expose some value cases as weak, and restrict some technically impressive capabilities. Those are not defects in the discipline. They are how capital, trust, and human attention are protected for the interventions that deserve them.

Such an organization will still make mistakes. AI Value Management is not a promise of perfect prediction. Its advantage is that errors remain observable and revisable. Weak assumptions can be reopened. Authority can move. A pilot can stop. A runtime incident can change permission. A portfolio can release capacity. The system is built to learn without surrendering accountability.

The organization that learns to stop well will often scale better. The organization that records exceptions can improve its standards. The organization that keeps local outcome owners visible can move faster without confusing delegation with abdication. The organization that can challenge a fluent answer from the wrong frame remains capable of thinking above the output it receives.

Begin with one real commitment

The smallest serious starting point is not an enterprise transformation programme. Choose one consequential AI use that the organization is considering, piloting, or already

operating.

Name the outcome and the realistic alternative. Draw the workflow as it actually operates. Identify the owner and the people who carry the benefit, burden, and downside. Define the system boundary and the actions it may take. Establish the evidence required for the next commitment. Record who may approve, challenge, correct, suspend, and change it. Decide when the authorization expires or reopens.

Then make the decision.

The result may be to proceed. It may be to narrow the boundary, repair the workflow and operating conditions around the AI, run a better experiment, build a shared capability, or stop. The value lies in replacing momentum with an explicit commitment that can be examined and changed.

Repeat that discipline and the organization begins to change. Portfolio claims become comparable. Shared constraints become visible. Governance becomes infrastructure. Expertise is placed where it matters. Evidence alters funding and permission. Management becomes capable of directing AI rather than merely reacting to it.

■ IN PRACTICE

Choose one consequential AI use and complete the seven steps in this paragraph within a week. Then make the decision.

The organization must be able to think above the output

That is the argument of this book. Management is not a passive recipient of AI capability waiting to discover what the technology will do to the organization. Management shapes what AI becomes here by designing the conditions under which the capability is allowed to matter.

The enduring human role is sometimes described as oversight, but that word can be too passive. Oversight suggests that the important work has already been performed by the system and that a person merely checks the result. The deeper responsibility is judgment: deciding what deserves to be asked, what standard should govern the answer, what is missing from the frame, and what consequence the organization is willing to own.

That responsibility becomes more important when the output looks intelligent. A visibly broken answer invites correction. A polished answer can travel through a meeting, a report, a workflow, or an executive decision without creating the feeling that anything has failed. The organization therefore needs people and structures capable of challenging plausible work before plausibility becomes authority.

This is not an argument for keeping humans in every loop forever. Some decisions can and should become more automated as the evidence, boundary, reversibility, and consequence justify it. The point is that the organization must retain the practical ability

to redefine the problem, change the rule, inspect the evidence, provide remedy, and withdraw permission.

AI can make expertise more consequential because one person with strong judgment can direct enormous capability. It can also make the absence of expertise harder to detect. Management has to design for both realities.

The question is therefore larger than whether the AI works. The question is whether the organization remains capable of understanding what the AI is doing to the work, the decision, the people affected, and the value it claims to create.

Close the gap

AI can answer the question an organization asks. It cannot guarantee that the organization asked from the right level of reality.

It cannot decide whose outcome matters, which assumption deserves challenge, what consequence is acceptable, when a plausible answer is insufficient, or whether a result was worth creating. It cannot guarantee that the person asked to review it knows what good looks like or has the authority to act on what they see.

That responsibility remains human. More precisely, it remains organizational. It belongs in how problems are framed, work is designed, authority is allocated, evidence is preserved, trade-offs are made, and learning returns to the system.

The future of AI inside an organization is not something management simply discovers after deployment. It is something management shapes through the conditions it builds and the decisions it is willing to own.

This book is an invitation to build that organization.

An organization that does not confuse adoption with value, fluency with judgment, or technical permission with permanent legitimacy. An organization capable of moving quickly without abandoning consequence, learning without pretending uncertainty is proof, and using machine capability without surrendering command of the outcome.

AI capability can be purchased.

The management capacity required to turn it into worthwhile value must be built.

Close the gap.

How AI Value Management Relates to Existing AI Governance and Management Systems

AI Value Management works alongside established risk, governance, quality, project, benefits-realization, and AI management systems. Its job is to keep those specialist contributions connected around one bounded AI-enabled intervention, one value case, one operating boundary, and one next management decision.

NIST AI Risk Management Framework

The NIST AI Risk Management Framework organizes AI risk work through Govern, Map, Measure, and Manage.¹ Those functions are compatible with AI Value Management but operate at a different level of abstraction. NIST helps an organization establish governance, understand context and risk, measure relevant characteristics, and manage identified risk. AI Value Management uses those contributions inside the broader question of whether a specific intervention should be selected, placed, authorized, tested, operated, expanded, restricted, or stopped.

For example, mapping the context and affected stakeholders can strengthen the intervention boundary. Measurement and monitoring can supply evidence for readiness and runtime assurance. Governance roles and risk tolerances can constrain decision rights and operating rules. None of those steps, by themselves, establish that the intervention creates worthwhile net value or that the next commitment is justified.

ISO/IEC 42001 and AI management systems

ISO/IEC 42001 provides a management-system structure for establishing responsibilities, policies, objectives, processes, monitoring, evaluation, and continual improvement around organizational AI.² AI Value Management can operate inside that system. The Five Conditions do not compete with an AI management system; they provide an intervention-level management lens that asks whether value clarity, workflow fit, operating rules, ownership/accountability, and evidence/learning are coherent enough for a particular commitment.

An organization using ISO/IEC 42001 may already possess useful artifacts: role definitions, risk processes, documented controls, monitoring practices, change procedures, and management-review mechanisms. The toolkit should reuse those records where they answer the intervention-level question rather than duplicate them.

Benefits realization, investment appraisal, and portfolio management

Benefits-realization and investment-appraisal disciplines contribute the baseline, alternative, value mechanism, cost, timing, ownership, and benefit-tracking logic used in the Value Case.³ Portfolio management contributes the resource-allocation discipline used in Chapter 15. AI Value Management adds a specific constraint: technical performance, adoption, and local task gains cannot be allowed to stand in for realized organizational consequence.

Risk and governance remain part of value

The book deliberately does not treat risk as a separate veto layer applied after the value case. Review burden, remedy, privacy, safety, workforce consequence, legal constraint, and operational resilience can change the economics and legitimacy of the intervention itself. A use that creates gross efficiency while shifting unacceptable downside elsewhere has not produced the value management claimed.

The compatibility rule

Use existing systems wherever they already produce a decision-useful artifact. Do not create a second control merely to satisfy AI Value Management terminology. Add a new record only when a material management question remains unanswered.

The practical test is simple: can management reconstruct the intervention boundary, intended outcome, workflow placement, operating rules, accountable owner, evidence state, critical limitations, and the decision currently authorized? If existing governance and management systems can answer those questions, they should be reused. If they cannot, the gap is the work.

This appendix describes conceptual compatibility, not formal clause-by-clause equivalence, certification, or compliance assurance. Any regulatory, legal, safety, security, privacy, employment, or professional obligations remain subject to the relevant specialist authority.

Glossary

AI Value Management	The discipline of selecting, designing, authorizing, operating, proving, and adapting specific business uses of AI so that capability becomes worthwhile, measurable, and sustainable value.
AI Value Office	The compact coordinating function that keeps the chain visible. It owns the management system; business owners own the results.
AI-enabled intervention	One defined business use of AI inside a defined system of work: its workflow, people, data, decisions, permissions, costs, controls, stakeholders, evidence, and current authorization.
Denominator	Cost per token describes consumption. Cost per accepted output or completed case is closer to the work. Cost per realized outcome is closer to the consequence.
Engineered Epistemic Performance	Knowledge-shaped language, behaviour, and interface signals that lead people to attribute more understanding, completeness, or authority to a system than the evidence warrants.
Evidence grade	Assumption, documented design, controlled evaluation, shadow operation, bounded live evidence, or sustained operation. Confidence and dissent are recorded separately.
Five classifications	Utility, experiment, strategic candidate, watch, and reject. All five are legitimate. Confusing them misallocates capital.
Five uncertainties	Capability, workflow, human behaviour, consequence and control, and economics and scaling. A pilot is designed around the one blocking the next decision.
Four pathways	Self-service, delegated team, specialist, and executive. Place each decision on the least burdensome route that remains credible for its consequence.

Friction and function	A handoff that only transfers or re-enters information is friction. A handoff that carries judgment, accountability, learning, trust, or remedy is function.
Governance measures	Discovery coverage, time to decision, share on standard pathways, evidence reused, specialist queue age, remediation time, and decisions reversed.
Level of the claim	Task, workflow, business unit, or enterprise. A gain is real at the level it was measured. It becomes a different claim at the next level.
Performance floor	The quality, reliability, latency, and review burden this task and consequence require – strong enough to protect the outcome, narrow enough to avoid buying unneeded capability.
Portfolio classes	Current value, strategic experiment, and shared foundation. Compare within a class; allocate across classes by strategy and scarce capacity.
Portfolio tax	The duplicated uncertainty created when every team independently chooses vendors, negotiates data access, invents review rules, and requests bespoke approval.
Proportional evidence	Enough evidence for the next reversible step; stronger evidence before a choice creates durable structure, obligation, exposure, or dependence.
Proportional federation	Central capability builds the system; local owners control outcomes; embedded specialists connect the two; independent challenge can say no, not yet, or only inside this boundary.
Realized benefit	A gain that has changed the operating or financial system – not potential capacity still waiting to be absorbed.
Review modes	Observe, confirm, decide, and dual control. Match the mode to consequence, uncertainty, reversibility, detectability, and volume.
Runtime assurance	Observation connected to interpretation, decision, intervention, revalidation, and an updated authorization or exit.

The reversed sequence	Outcome or uncertainty → workflow and people → where AI enters → data, authority, review, and exit → only then the tool.
Three evidence loops	Learning before structural deployment, operating after deployment, and value and impact across both. They must meet in a forum with authority.
Value case	A living explanation of an intervention's outcome, comparison, mechanism, alternatives, full cost, evidence, owner, and next decision.

Notes

INTRODUCTION

1. BlackRock Investment Institute. “2026 Midyear Global Investment Outlook.” June 2026.
2. Stanford HAI. The 2026 AI Index Report. 2026.
3. Boston Consulting Group. “As AI Investments Surge, CEOs Take the Lead.” BCG AI Radar 2026.
4. Innovation, Science and Economic Development Canada. “Canada’s National Artificial Intelligence Strategy: AI for All.” 2026.
5. PwC. 29th Global CEO Survey. 2026.
6. Deloitte. The State of AI in the Enterprise 2026.
7. McKinsey & Company. “From Promise to Impact: How Companies Can Measure and Realize the Full Value of AI.”
8. Scalebotics Research. The State of AI Value 2026 (200-company screen; 300 claim hypotheses; 220 selected public-evidence profiles).
9. Noy, Shakked, and Whitney Zhang. “Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence.”; Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. “Generative AI at Work.”
10. Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. “Artificial Intelligence and the Modern Productivity Paradox.”; Brynjolfsson, Rock, and Syverson. “The Productivity J-Curve: How Intangibles Complement General Purpose Technologies.”

CHAPTER 1: ADOPTION IS NOT VALUE

1. Klarna. “AI Assistant Handles Two-Thirds of Customer Service Chats in Its First Month.” February 2024.
2. Reuters. “Sweden’s Klarna Shifts AI Focus from Cost Cuts to Growth.” September 2025.
3. Scalebotics Research. The State of AI Value 2026. Counts reproduced from the 220-claim benchmark set.
4. Dillon, Eleanor Wiske, Sonia Jaffe, Nicole Immorlica, and Christopher T. Stanton. “Shifting Work Patterns with Generative AI.” NBER Working Paper 33795.

FURTHER READING

- MIT Sloan. “Metrics of Meaning: What Do We Really Measure in AI?”
- McKinsey & Company. “From Promise to Impact: How Companies Can Measure and Realize the Full Value of AI.”
- U.S. Government Accountability Office. “IT Investment Management: SSA Needs to Oversee Investments in Operations and Better Evaluate Performance.”

CHAPTER 2: CAPABILITY IS NOT STRATEGY

1. OpenAI. “Morgan Stanley Uses AI Evals to Shape the Future of Financial Services.”

2. Morgan Stanley Wealth Management. “Launch of AI @ Morgan Stanley Debrief.” 26 June 2024.
3. Porter, Michael E. “What Is Strategy?” *Harvard Business Review* 74, no. 6 (1996): 61–78.

FURTHER READING

- Montard, Antoine, Dago Diedrich, and Tanguy Catlin. “Where AI Will Create Value – and Where It Will Not.” McKinsey, 29 April 2026.
- Diedrich, Dago, et al. “From AI Table Stakes to AI Advantage: Building Competitive Moats.” McKinsey, 15 May 2026.
- Demirer, Mert, et al. “Chaining Tasks, Redefining Work: A Theory of AI Automation.” NBER Working Paper 34859, February 2026.
- Teece, David J., Gary Pisano, and Amy Shuen. “Dynamic Capabilities and Strategic Management.” *Strategic Management Journal* 18, no. 7 (1997): 509–533.
- Barney, Jay B. “Firm Resources and Sustained Competitive Advantage.” *Journal of Management* 17, no. 1 (1991): 99–120.
- Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. *Power and Prediction*. Boston: Harvard Business Review Press, 2022.

CHAPTER 3: THE TOOL-FIRST TRAP

1. British Columbia Civil Resolution Tribunal. *Moffatt v. Air Canada*, 2024 BCCRT 149.

FURTHER READING

- Microsoft WorkLab. 2026 Work Trend Index Annual Report: Agents, Human Agency, and the Opportunity for Every Organization.
- KPMG. “Why Enterprise AI Maturity Stalls After Pilot Success.”
- KPMG. “AI-Ready Data: Close the Five Gaps Preventing Enterprise AI from Scaling.”
- ISACA. “Cutting Through the AI Hype: The 2026 AI Pulse Poll.”
- MIT Industrial Performance Center. “Humans in the Loop: The Evolution of Work in Early Experiments with Generative AI.”
- InformationWeek. “Tips for Successfully Exiting AI Vendor Contracts.”
- Markus, M. Lynne, and Robert I. Benjamin. “The Magic Bullet Theory in IT-Enabled Transformation.”
- Orlikowski, Wanda J. “The Duality of Technology: Rethinking the Concept of Technology in Organizations.”

CHAPTER 4: AI SCALES THE ORGANIZATION YOU ALREADY HAVE

1. Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.”
2. Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt. “Information Technology, Workplace Organization and the Demand for Skilled Labor: Firm-Level Evidence.”; Brynjolfsson, Erik, and Lorin M. Hitt. “Beyond Computation: Information Technology, Organizational Transformation and Business Performance.”

3. Brynjolfsson, Li, and Raymond. “Generative AI at Work.”
4. Parasuraman, Raja, and Victor Riley. “Humans and Automation: Use, Misuse, Disuse, Abuse.” *Human Factors* 39, no. 2 (1997): 230–253.
5. Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” FAccT 2021.
6. Li, Zhang, and Fan. “Understanding the Influence of Perceived Intelligence and Perceived Anthropomorphism on Users’ Intention to Adopt Healthcare Chatbots.” *Digital Health* (2026).
7. Reuters. “Mark Zuckerberg Had a Bold Plan to Replace Meta Staff with AI. Here’s How It Imploded.” 26 August 2026.

FURTHER READING

- Deloitte. *The State of AI in the Enterprise 2026*.
- IBM Institute for Business Value. *2026 CEO Study: Rewiring the C-Suite*.
- Harvard Business School AI Institute. “The Cybernetic Teammate: How AI Is Reshaping Collaboration and Expertise in the Workplace.”
- MIT Sloan. “How AI Is Reshaping Workflows and Redefining Jobs.”
- Bloom, Nicholas, Raffaella Sadun, and John Van Reenen. “Management as a Technology?”
- U.S. Equal Employment Opportunity Commission. “iTutorGroup to Pay \$365,000 to Settle EEOC Discriminatory Hiring Suit.”

CHAPTER 5: AI VALUE MANAGEMENT AND THE FIVE CONDITIONS

1. Project Management Institute. *Benefits Realization Management Framework*; National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; International Organization for Standardization. *ISO/IEC 42001:2023 – AI Management Systems*; *ISO/IEC 42005:2025 – AI System Impact Assessment*.
2. The billing-support assistant is an illustrative composite used in Chapters 5 to 7; it does not describe a specific company.
3. Demirer, Mert, et al. “Chaining Tasks, Redefining Work: A Theory of AI Automation.” NBER Working Paper 34859, February 2026.

FURTHER READING

- National Institute of Standards and Technology. *Challenges to the Monitoring of Deployed AI Systems (NIST AI 800-4)*.
- McKinsey & Company. “From Promise to Impact: How Companies Can Measure and Realize the Full Value of AI.”
- FinOps Foundation. “AI Value – FinOps Topic.”

CHAPTER 6: DEFINE THE VALUE CASE BEFORE THE USE CASE

1. HM Treasury. *Magenta Book 2026: Central Government Guidance on Evaluation*.

2. HM Treasury. The Green Book 2026: Central Government Guidance on Appraisal and Evaluation.

FURTHER READING

- Infrastructure and Projects Authority. Guide for Effective Benefits Management in Major Projects.
- Project Management Institute. PMI Lexicon of Project Management Terms.
- FinOps Foundation. “Managing AI Value Using FinOps Practice Operations.”
- RAND Corporation. The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed.

CHAPTER 7: FROM OUTPUT TO OUTCOME

1. Yotzov, Ivan, Jose Maria Barrero, Nicholas Bloom, Steven J. Davis, Brent Meyer, and Emil Mihaylov. “Firm Data on AI.” NBER Working Paper 34836, 2026.
2. Dillon, Jaffe, Immorlica, and Stanton. “Shifting Work Patterns with Generative AI.” NBER Working Paper 33795, revised November 2025.
3. Scalebotics Research. The State of AI Value 2026 and AI Value Performance Benchmark methodology. Public-evidence dimensions assess the inspectability and decision-usefulness of selected public claims; they do not infer undisclosed internal company performance. Counts are from the 220-claim benchmark set; the wider-market sample (384 claims screened, 211 profiled) is reported separately in the same study.
4. Brynjolfsson, Li, and Raymond. “Generative AI at Work.” NBER Working Paper 31161.
5. Brynjolfsson, Rock, and Syverson. “The Productivity J-Curve: How Intangibles Complement General Purpose Technologies.”

FURTHER READING

- McKinsey & Company. “Putting AI to Work: The Operational Excellence Imperative.”
- FinOps Foundation. “Token Economics: The Atomic Unit of AI Value.”
- FinOps Foundation. “Capability: Unit Economics.”

CHAPTER 8: WHERE AI BELONGS IN THE WORKFLOW

1. Dell’Acqua, Fabrizio, et al. “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality.”
2. The manufacturing workflow-placement and maintenance scenes in Chapters 8 and 10 are illustrative composites; they do not describe a specific company.
3. Demirev et al. “Chaining Tasks, Redefining Work: A Theory of AI Automation.” NBER Working Paper.

FURTHER READING

- MIT Sloan. “How AI Is Reshaping Workflows and Redefining Jobs.”
- MIT Industrial Performance Center. “Humans in the Loop: The Evolution of Work in Early Experiments with Generative AI.”

- MIT CISR. “Leveraging Digital Colleagues for Enterprise Value.”
- NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.

CHAPTER 9: WHICH MODEL BELONGS IN THE WORK?

1. Department for Transport and The Alan Turing Institute. “AI Consultation Analysis Tool (CAT) v1.0 Evaluation.” 23 December 2025.
<https://www.gov.uk/government/publications/ai-consultation-analysis-tool-evaluation>.
Reported savings are evaluation estimates.
2. OpenAI. “Morgan Stanley Uses AI Evals to Shape the Future of Financial Services.”
<https://openai.com/index/morgan-stanley/>. First-party case material.
3. OpenAI. “Introducing GPT-5.4 Mini and Nano.” 17 March 2026.
<https://openai.com/index/introducing-gpt-5-4-mini-and-nano/>.

CHAPTER 10: DECISION RIGHTS, HUMAN JUDGMENT, AND ESCALATION

1. Federal Trade Commission. “Rite Aid Banned from Using AI Facial Recognition After FTC Says Retailer Deployed Technology without Reasonable Safeguards.”; Federal Trade Commission. Complaint for Permanent Injunction and Other Relief: FTC v. Rite Aid.

FURTHER READING

- MIT CISR. “Designing Decision Rights for AI” and “The AI Decision Matrix.”
- ISACA. “New ISACA Research: 56% of Digital Trust Pros Do Not Know How Fast They Could Shut Down AI After a Security Incident.”
- NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.
- National Transportation Safety Board. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018.
- NIST. Challenges to the Monitoring of Deployed AI Systems (NIST AI 800-4).
- Buçinca, Zana, Maja Barbara Malaya, and Krzysztof Z. Gajos. “To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making.”

CHAPTER 11: GOVERNANCE AS OPERATING INFRASTRUCTURE

1. Microsoft Digital. “Governing AI Agents at Scale: Lessons from Our Journey at Microsoft.”

FURTHER READING

- Deloitte AI Institute. The State of AI in the Enterprise 2026.
- NIST. Artificial Intelligence Risk Management Framework 1.0 and AI RMF Playbook – Govern.
- ISO. “ISO 42001 Explained.”
- Office of the Superintendent of Financial Institutions. Guideline E-23 – Model Risk Management.
- DBS Bank. “DBS Bank’s AI Playbook.”

CHAPTER 12: READY FOR WHAT, EXACTLY?

1. Beede, Emma, et al. “A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy.”
2. The card-protection readiness and runtime scenes in Chapters 12 and 14 are illustrative composites; they do not describe a specific bank.

FURTHER READING

- Lavin, Alexander, et al. “Technology Readiness Levels for Machine Learning Systems.”
- NIST. Artificial Intelligence Risk Management Framework 1.0, Map function.
- Tikhomirov, Lana, et al. “A Scoping Review of Silent Trials for Medical Artificial Intelligence.”
- U.S. Government Accountability Office. Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities.

CHAPTER 13: PILOTS THAT PRODUCE DECISIONS

1. UK Home Office. Evaluation of AI Trials in the Asylum Decision-Making Process.
2. The procurement pilot scene is an illustrative composite; its comparison, measures, and decision are examples rather than claims about a named organization.
3. Associated Press. “McDonald’s Is Ending Its Test Run of AI-Powered Drive-Thrus with IBM.” June 2024.

FURTHER READING

- HM Treasury. The Magenta Book: Central Government Guidance on Evaluation.
- HM Treasury / UK Government. Guidance on the Impact Evaluation of AI Interventions.
- NIST. Practices for Automated Benchmark Evaluations of Language Models (NIST AI 800-2, initial public draft).

CHAPTER 14: KEEP VALUE AND PERMISSION CURRENT

1. Zillow Group. Third-Quarter 2021 Results and plan to wind down Zillow Offers.
2. Department for Transport. “AI Consultation Analysis Tool v1.0 Evaluation.”

FURTHER READING

- NIST. Challenges to the Monitoring of Deployed AI Systems (NIST AI 800-4).
- NIST AI Resource Center. AI RMF Playbook – Manage.
- U.S. Food and Drug Administration. Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions.
- U.S. Government Accountability Office. Artificial Intelligence Acquisitions: Selected Practices and Lifecycle Considerations.

CHAPTER 15: MANAGING THE AI PORTFOLIO

1. U.S. Government Accountability Office. “Artificial Intelligence: IRS Actions Needed to Address Skills Gaps, Information Quality, and Strategic Management.” 2026.
2. Scalebotics Research. The State of AI Value 2026. Intervention and portfolio claims are reported in separate comparison panels; panel averages describe the selected public-

evidence sample.

3. The portfolio trade-off meeting and the commitments inside it are illustrative composites.

FURTHER READING

- U.S. Government Accountability Office. “Generative AI Use and Management at Federal Agencies.”
- NIST. AI Risk Management Framework Playbook – Govern (inventory, prioritization, and decommissioning).
- Project Management Institute. The Standard for Portfolio Management, Fourth Edition.

CHAPTER 16: BUILDING THE AI VALUE ORGANIZATION

1. The AI Value Office is a proposed coordinating function, not a claim that one reporting line or title fits every enterprise.
2. The first-ninety-days sequence is the author’s implementation proposal.

FURTHER READING

- Microsoft. 2026 Work Trend Index.
- World Economic Forum. 2026 workforce and organizational-design synthesis.

CONCLUSION

1. The forecasts in this chapter are the author’s strategic judgment: access to capable AI is expected to diffuse, increasing the relative importance of placement, evidence, authority, adaptation, and allocation.

APPENDIX A

1. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework and associated Playbook resources.
2. International Organization for Standardization. ISO/IEC 42001, Artificial Intelligence – Management System.
3. HM Treasury appraisal and evaluation guidance; Project Management Institute portfolio-management principles; benefits-realization guidance cited in Chapters 6, 14, and 15.

Index

A

Adoption-to-Value Ladder	22
AI Value Management	14, 56–59, 164, 185, 189, 192–193
AI Value Office	180
AI-enabled intervention	56, 167, 192
Air Canada	34
attribution	11, 20, 25, 47, 72–73, 85, 93, 133

B

baseline	20, 78, 151
BCG	10
bottleneck	45, 51, 62, 74, 88–89, 92–95, 97, 101, 104, 112, 132, 182
Brynjolfsson, Erik	88

C

capability frontier	101, 105
change contract	115, 161
complementary investment	12, 30–31, 43, 94
conditional readiness	144
counterfactual	71–73, 80, 172
critical gates	141–143

D

decision rights	43, 118–119, 133, 180, 192
Deloitte	11
denominator	76, 90, 112–113, 115
Department for Transport (UK)	110, 162
disbenefits	61, 71–72, 75, 77
dual control	121

E

Engineered Epistemic Performance	47–48, 125
escalation	12, 119, 122
evidence and learning (condition)	61, 65
evidence grade	142–143

G

Google Health	139
---------------	-----

governance debt	135
Government Accountability Office (US)	169

H

HM Treasury	74, 81
Home Office (UK)	151
hours saved	23, 92
human in the loop	64, 118, 121

I

ISO/IEC 42001	192
---------------	-----

K

Klarna	19
--------	----

M

Managed Pilot Contract	150
McDonald's	154
McKinsey	11
Meta, Project OT	49, 91
Microsoft 365 Copilot	23, 84
Microsoft Digital	129
Morgan Stanley	27, 30, 113

N

NIST AI Risk Management Framework	192
-----------------------------------	-----

O

operating rules (condition)	58, 60, 63, 67–68, 192–193
option value	13, 62, 80
Output-to-Outcome Chain	86–87, 95, 101
ownership and accountability (condition)	61, 64

P

performance floor	30, 107, 110–111, 114–115
Porter, Michael	28
portfolio tax	36
proportional federation	184
proportionality	67, 72, 81, 86, 90, 131–132, 164, 184
PwC	11

R

review	35, 43, 105, 124, 136, 140, 160, 170–171
review modes	121
Rite Aid	118
routing	60, 95, 105, 111–112, 114, 123, 149, 153, 174
Runtime Assurance Loop	159

S

safe state	121–123, 126, 162
Scalebotics AI Value	12, 22, 85, 172
Performance Benchmark	
shadow AI	38, 64, 129, 135, 140
Stanford AI Index	10

Strategic AI Test	29
-------------------	----

T

tool-first trap	32, 34, 39
-----------------	------------

V

value case	14, 72, 76–77, 80, 163
value clarity	58, 60–62, 67–68, 71, 192
(condition)	
value mechanism	74, 78, 104, 193

W

workflow fit	23, 36, 58, 60, 62–63, 67–68, 143, 149, 192
(condition)	

Z

Zillow	158
--------	-----

About the Author

Layefa Okiy is the founder of Scalebotics and leads the company's research, doctrine, judgment, and product direction. The broader author, speaking, and research presence is at layefaokiy.com.

About Scalebotics

Scalebotics is a services-led, software-enabled, research- and IP-driven company. It connects enterprise direction, portfolio choices, defined AI use design, operating authority, deployment, and evidence of realized outcomes.

The AI Value Gap is the management doctrine behind Scalebotics' research, services, and software. The companion operating toolkit and current research are at scalebotics.io.

