PART I · THE ARGUMENT

A Science and Its Discipline

OPENING CASE

The discipline that changed its mission

DATED: FEBRUARY 2026. FIGURES AND THE QUOTATION CHECKED AGAINST THE FOUNDATION'S REPORT AND ANNOUNCEMENT, OCTOBER 2026.

In February 2026, the FinOps Foundation changed its mission statement. For years, it had described its purpose as "advancing the people who manage the value of cloud." The new statement replaced one word: "advancing the people who manage the value of technology."[1]

The change accompanied the Foundation's annual survey of its practitioners, the people inside companies who manage cloud spending for a living. The 2026 survey reached 1,192 respondents responsible for more than $83 billion in annual cloud spending.[2] Its central finding was how far FinOps had expanded into AI spending: 98 percent of respondents now managed it, up from 31 percent two years earlier.[3]

FinOps was built to address a specific problem: cloud bills varied with use and arrived after capacity had been consumed.[4] The discipline developed to make engineering, finance, and business teams jointly accountable for that spending.[5] Its practitioners later took on AI spend-

1. FinOps Foundation, "A One Word Change: How the FinOps Community Made Our Mission Evolution Inevitable," February 19, 2026, accessed October 17, 2026. The mission statement before and after the change.

2. FinOps Foundation, "State of FinOps 2026," February 19, 2026, accessed October 17, 2026. Survey size and the cloud spending its respondents manage.

3. FinOps Foundation, "State of FinOps 2026," February 19, 2026, accessed October 17, 2026. Share of respondents managing AI spend, 2026 survey against 2024 survey.

4. FinOps Foundation, "2024 Changes to the Definition of Cloud FinOps," accessed October 17, 2026. Cloud consumption as variable and tracked after the fact.

5. FinOps Foundation, "What is FinOps?," accessed October 17, 2026. The Foundation's definition.

ing, and by 2026 nearly all respondents to the Foundation's survey were managing it.

Asked whether AI was providing value, one practitioner acknowledged the uncertainty. The report quoted the answer: "No one can answer that question yet."[6]

The quotation expresses the gap Chapters 1 and 2 described, in the words of someone asked to close it. When an organization pays for a metered resource, costs accrue with use, but someone must build a way to measure the value that use returns. FinOps is a capable neighboring discipline that has expanded into work it was not built to govern: the economics of an organization's AI consumption.

That work requires both a body of knowledge about what is true and a practice that turns that knowledge into action. This book sets out both. Chapter 14 explains how FinOps and that practice divide the work. Until both exist, the practitioner's question about whether AI is providing value remains unanswered.

3.1 The science and the practice

Managers act on claims about how their businesses work. A budget assumes that costs follow usage. A pricing negotiation assumes that the provider has a reason to accept a lower price. A dashboard assumes that the information it displays will change someone's decisions. When those claims are wrong, careful execution cannot repair the decisions built on them.

Claims like these can remain informal, drawn from experience, rules of thumb, or the last presentation's slides. That knowledge is useful, but leaving it unstated has a cost: two managers can hold opposing beliefs about the same deployment without being able to identify where they disagree. One might treat Chapter 1's Cursor licenses as a fixed software

6. FinOps Foundation, "State of FinOps 2026," February 19, 2026, accessed October 17, 2026. Practitioner quotation as reproduced in the report.

expense and the other as metered consumption. Unless they state their claims, neither can identify the fact that would settle the disagreement, so the argument turns on confidence rather than evidence.

Chapters 1 and 2 each ended with an explicit result rather than an opinion. Chapter 1 stated the conditions under which deployed AI is a resource the organization consumes. Chapter 2 stated the conditions under which a scaled deployment cannot be controlled without a governing apparatus.

AI Business Economics is that body of knowledge. It is a science in a specific, limited sense: a set of claims about the business economics of consuming AI, each stated precisely enough to be checked. Here, science does not imply laboratory experiments or certainty. It means that every claim states when it applies, what it concludes, and which other claims it rests on.

DEFINITION

AI Business Economics

The science of the business economics of consuming AI: an ordered body of conditional claims, each stating when it applies, what it concludes, and which other claims it depends on.

Stating when a claim applies makes it useful to a manager. Chapter 1's result does not say that every use of AI is consumption. It identifies the conditions under which deployed AI behaves as a resource the organization consumes, so a manager can check a deployment against those conditions before relying on the result. Stating those conditions also makes it possible to show when the claim does not apply. Two managers who disagree can then identify the fact that would settle the question.

Chapter 1's Cursor licenses show this check in practice. Two of the result's conditions are that the buyer's cost extends beyond the access price and that measurement makes resource use visible. From June 16, 2025, the individual Pro plan included twenty dollars a month of frontier-model use at the model provider's rates, and further use was priced at the same rates. For a developer whose frontier-model use ran past

that credit, both conditions held.[7] A manager who still budgeted the licenses as a fixed software expense was treating the access price as the full cost, even though the conditions showed otherwise. Checking those conditions could settle what confidence alone could not.

The book presents the results of this science in numbered panels, as Chapters 1 and 2 did. Each panel states a result in its exact form and carries an identifier, followed by an explanation in plain business language. The panel and its explanation give a manager enough to use the result. Readers who want to examine it further can use the identifier to find its full statement and supporting claims in Appendix A.

Chapter 2's panel, THM-004, shows what such a result settles and what it leaves open. It formalizes the economic consequence of the third flow's asymmetry. Once a deployment reaches scale, economic control requires a governing apparatus. Diligence from the people running the deployment is no longer enough. The theorem does not establish the three flows themselves. They are this book's way of examining a deployment, and the theorem's conditions do not mention them.

3.2 The five functions

A manager can know that consumption drives cost and still have no forecast of next month's use, no owner for the bill, and no record of which team used what. Knowledge alone does not change what an organization does. A practice connects knowledge to results through methods, roles, and standards, as clinical practice does for physiology and engineering codes do for physics.

Chapter 1 described the problem as a category error. Wherever that chapter's conditions hold, AI Operations Management addresses the error by managing AI as a resource the organization consumes. The discipline has five functions that follow AI capacity through the organization:

---

7. Michael Truell, "Clarifying Our Pricing," *Cursor*, July 4, 2025, accessed October 17, 2026. The Cursor post describing the plan change of June 16, 2025, as cited in Chapter 1.

buying it, planning its use, recording and assigning consumption, routing it to work, and judging the value it returns. Each function is a capability an organization either has or lacks. Without one, the organization continues consuming but cannot answer the question that function addresses.

The first function, sourcing, matches the AI capacity an organization buys to the requirements of its work. It determines which models and providers serve which work, and on what terms, so the organization neither pays for capability it does not need nor deprives work of capability it requires. A company that sends every request to its most capable model, including requests a smaller model would answer just as well, is paying for capability that work does not need.

Sourcing involves more than choosing a provider. Requirements may include accuracy, speed, cost per request, and ease of connection to the systems that feed a model. Different kinds of work place different weight on each requirement: customer messages may need a fast, inexpensive model, while a contract review may need the most capable one available. An organization that buys one model for all its work has still made a sourcing decision, whether or not anyone made it deliberately.

DEFINITION

AI Operations Management

The management discipline that acts on AI Business Economics: the practice by which an organization buys AI capacity, plans its use, records and assigns consumption, allocates capacity, and holds that consumption to account.

> **DATED: OCTOBER 2025**
>
> In October 2025, Airbnb's chief executive, Brian Chesky, said the company's customer-service agent was built on 13 models from OpenAI, Google, Alibaba, and open-source providers. He said the company relied heavily on Alibaba's Qwen models, which he called "very good" and "also fast and cheap." He also said OpenAI's tools for making Airbnb's app available inside ChatGPT were not yet ready.[8]

The second function is planning and budgeting, whose absence proved costly in Chapter 2's opening case. The organization must state how much AI it expects to consume in the coming period and detect when actual consumption departs from that expectation. A budget is a commitment made in advance. Comparing it with recorded consumption throughout the period makes deviations visible before the money runs out.

An AI budget differs from a license budget because the quantity consumed is not fixed when the budget is set. The plan must therefore state its assumptions about what drives consumption, including how many people or workflows will use the system and how heavily. Managers can then trace a change in spending to a change in one of those drivers while there is still time to act.

> **DATED: OCTOBER 2026**
>
> Pigment, a company that sells financial planning software, surveyed about 2,000 finance leaders in the United States, the United Kingdom, France, and Germany for a report released in October 2026. Eighty-three percent said consumption-based pricing had pushed their AI spending above expectations. Thirteen percent said they had stayed within budget, and the report put the average overrun at about 27 percent. These figures describe the survey's respondents and were reported by a vendor of planning software.[9]

---

8. "Chesky says OpenAI tools not ready for ChatGPT tie-up with Airbnb app," *Fortune*, October 21, 2025, accessed October 17, 2026. The chief executive's account of the models behind the agent.

9. Pigment, "CFO Index, Q3 2026," October 7, 2026, accessed October 17, 2026. Survey size, overruns, and the share within budget.

The third function, metering and attribution, builds what Chapter 2 called the record flow. The organization must record what was consumed and assign each unit of consumption to the team, product, or purpose it served. Without that assignment, managers can see a cost but cannot trace it to the work that caused it.

The record has two parts: metering counts what was consumed, request by request, while attribution assigns that consumption to the work that caused it. An invoice priced by use can supply an aggregate count; an invoice that reports only seats and charges supplies neither consumption counts nor attribution. Only the organization can attribute consumption, because only it knows which team, product, or customer a request served. Without that record, no one can say whose work caused a cost, and responsibility can remain divided among several groups.

> **DATED: FEBRUARY 2026**
>
> DoiT, a company that sells cloud cost-management services, commissioned a survey of 500 finance leaders at organizations with at least 1,000 employees in the United States and the United Kingdom. Sapio Research conducted it in February 2026. Seventy-nine percent reported AI-related cost overruns in the twelve months before the survey. Asked who was accountable for AI spending, 55 percent named technology leadership and 53 percent named finance.[10]

> **DATED: MAY 2026**
>
> In May 2026, ServiceNow said it had already used up its full-year budget for Anthropic's AI tools. Its chief digital information officer, Kellie Romack, said the provider did not automatically show customers which of their users consumed which tools, or how much. ServiceNow tracked its employees' daily use of the tools through its own monitoring application and assigned a member of her team to watch the data. "I want more transparency," she said.[11]

---

10. Isaac Hubley, "The AI Spending Data Your Board Is About to Ask You About," *DoiT blog*, June 9, 2026, accessed October 17, 2026. Survey size, overruns, and the split of accountability.

11. Laura Bratton, "Why Anthropic Costs Are Unpredictable," *The Information*, May 14, 2026, accessed October 19, 2026. The budget, the missing usage detail, the monitoring, and the quotation.

The fourth function is allocation and routing. When capacity is limited, the organization must decide what runs first and send each piece of work to suitable capacity. These choices occur even when no one makes them deliberately. Consider a constructed example: two workflows share a provider account with a fixed rate limit. When their combined demand exceeds the limit, one must wait, even though no one has decided which should take priority. Allocation and routing make that choice deliberate by applying a rule.

> **DATED: AUGUST TO SEPTEMBER 2026**
>
> In August 2026, a vice president of AT&T, Mark Austin, said the company's internal AI service, Ask AT&T, processed about 45 billion tokens a day and that open models powered 40 percent of employees' AI queries. He said AT&T planned to keep its employees' spending on Anthropic and OpenAI models flat in the coming years and that developers' spending was capped. He said that after AT&T began using LiteLLM to route employees' queries to cheaper models based on task complexity, the cost of some advanced AI tasks, including coding, fell by as much as 56 percent, while AI performance fell 2 percent.[12]
>
> In September 2026, the Financial Times reported that AT&T was using open models tuned on its own data to perform specific tasks at lower cost, and the company's chief data and AI officer, Andy Markus, said AT&T aimed to run 70 percent of its AI workloads on open models within a year.[13]

Routing and allocation are two sides of one decision. Routing sends each piece of work to suitable capacity, such as a smaller model for routine requests and a larger one for difficult cases. Allocation decides who goes first when capacity is insufficient for everyone. Both require a stated rule that managers can check against the results and change.

---

12. Aaron Holmes, "AT&T is Using Open Source Models to Curb Anthropic Bills," *The Information*, August 20, 2026, accessed October 19, 2026; Andy Markus, "The tokenomics equation," *AT&T*, July 23, 2026, accessed October 19, 2026. The vice president's account; the daily token figure also appears in the company's own post.

13. Ryan McMorrow and George Hammond, "Corporate America embraces cheaper 'open' AI models," *Financial Times*, September 27, 2026, accessed October 19, 2026. The tuned models and the target, as reported.

> **DATED: JULY 2026**
>
> In July 2026, Bloomberg reported that Microsoft had begun sending some prompts from Copilot in Excel and Outlook to its own MAI models instead of models from OpenAI and Anthropic, citing the cost of the outside models. Bloomberg put the volume at tens of thousands of prompts a week. The reporting did not say how the tasks that moved were chosen. The company's head of AI, Mustafa Suleyman, said its goal was to reduce and ultimately eliminate its payments to Anthropic.[14]

The fifth function is holding value to account. Consider a constructed example in which a support team adds an AI assistant to its help desk. At the end of a quarter, the organization should be able to say whether the assistant paid for itself, identifying its costs and returns, the months covered, and the outcome used to judge it. That requires a boundary specifying the costs and value the evaluation counts, the period it covers, and the outcomes against which it measures performance. The organization sets this value boundary for each deployment before spending begins and names one person responsible for knowing the answer.

The boundary matters because the choice of what to count changes the reported return. An evaluation that counts the assistant's subscription but excludes staff time spent checking its answers reports a higher return than one that counts both. An evaluation that measures only the assistant's best month rather than the whole quarter can also report a higher return. Setting the boundary in advance prevents those conducting the evaluation from choosing the version they prefer.

14. "Microsoft begins replacing third-party AI models with its own in Office apps," *Redmond Magazine*, July 9, 2026, accessed October 17, 2026. The routing change, its scale, and the stated cost motive.

DATED: OCTOBER 2025 AND OCTOBER 2026

In October 2025, JPMorgan Chase's chief executive, Jamie Dimon, told Bloomberg Television that the bank spent about $2 billion a year on AI and received about the same in benefits: "We have shown that for $2 billion of expense, we have about $2 billion of benefit."[15] In October 2026, the bank's chief financial officer, Jeremy Barnum, put annual AI spending closer to $700 million, while calling that estimate "a little stale at this point." He said the benefits were not yet observable in the bank's quarterly earnings. The reports give annual figures but do not say which costs and benefits each figure includes.[16]

Figure 3.1 shows the relationship between the two layers: the science provides the foundation for the discipline's five functions.

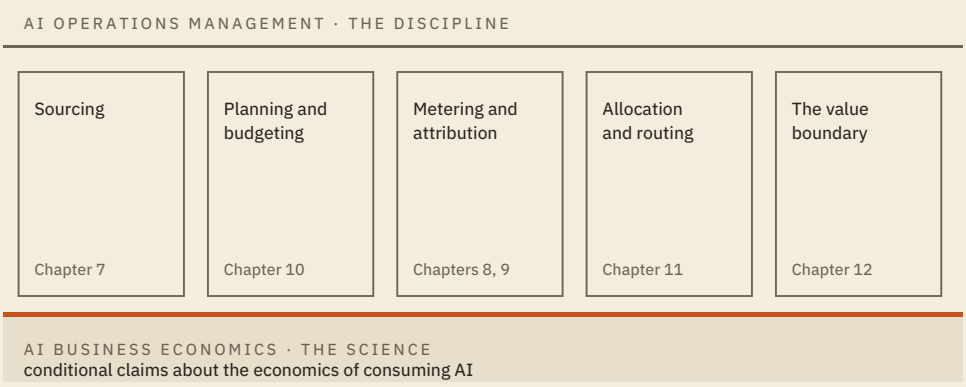


Figure 3.1 *The two layers. AI Business Economics states conditional claims about consuming AI and records what each one rests on. AI Operations Management acts on it through five functions, all taught in Part III.*

A result about governance changes nothing until an organization builds the apparatus it describes. The science therefore depends on the discipline to put its findings into practice. The discipline depends on the

15. "JPMorgan's Dimon says AI cost savings now match money spent," *Bloomberg News*, October 7, 2025, accessed October 17, 2026. The chief executive's figures.

16. Caitlin Mullen, "JPMorgan CEO Dimon warns customers of 'liability transfer' with AI," *Banking Dive*, October 13, 2026, accessed October 17, 2026. The later spending figure and the earnings observation.

science because careful practice cannot repair a decision built on an incorrect claim.

3.3 Where the discipline ends

Chapter 1 set aside three related subjects: how models work, prompt engineering, and finding uses for AI. Each has its own purpose and literature, and each ends before AI Operations Management begins.

The first subject concerns how models work, including their architectures, training, and internal behavior. It belongs to computer science and machine learning engineering and ends where a model becomes a service an organization can call. AI Operations Management begins when the organization can buy and call that service.

The second is prompt engineering, the craft of writing inputs that produce better outputs. It improves individual uses of AI, while AI Operations Management governs their combined consumption, cost, and value. A better prompt can lower the cost of one request. Determining whether the organization's requests, taken together, are worth what they cost requires the discipline.

The third subject is finding uses for AI: deciding which business problems it might solve. That work belongs to strategy and innovation and ends when a use is chosen. AI Operations Management begins with buying capacity for the chosen use and manages its consumption from then on.

Four established disciplines overlap more directly with AI Operations Management. Each manages an aspect of work related to AI consumption, but each lacks something needed to govern its economics. Table 3.1 compares them.

NEIGHBOR	WHAT IT IS FOR	WHAT IT MANAGES	WHAT IT LACKS FOR THIS TERRITORY
AIOps	Applying big data and machine learning to automate IT operations, including event correlation, anomaly detection, and causality determination[17]	The health and performance of IT systems	It uses AI to watch systems. It does not govern what consuming AI costs or returns
MLOps	Building, deploying, and monitoring machine learning models in production[18]	The lifecycle of models an organization builds	It ships models. It does not govern the consumption of capacity an organization buys
FinOps	Maximizing the business value of technology through financial accountability and collaboration among engineering, finance, and business teams[19]	Cloud and wider technology spending	It grew from managing cloud cost and usage. Its 2026 report quotes one practitioner saying that no one can yet answer whether AI is providing value[20]
Regulatory AI governance	Ensuring that uses of AI are lawful, safe, and accountable to regulators	Risk and compliance	It asks whether a use is permitted, not whether it is economically managed

Table 3.1. The four neighboring disciplines. Each manages work related to AI consumption, but none was built to govern its economics.

FinOps is the closest neighboring discipline, making its boundary with AI Operations Management the hardest to define. It already manages spending, has expanded into AI, and defines its purpose in terms of

17. Gartner, "AIOps (Artificial Intelligence for IT Operations)," *Gartner Information Technology Glossary*, accessed October 17, 2026. Gartner's definition of AIOps.

18. Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl, "Machine Learning Operations (MLOps): Overview, Definition, and Architecture," *IEEE Access* 11 (2023): 31866-31879. Definition and scope of MLOps.

19. FinOps Foundation, "What is FinOps?," accessed October 17, 2026. The Foundation's own definition of the practice.

20. FinOps Foundation, "State of FinOps 2026," February 19, 2026, accessed October 17, 2026. Practitioner quotation as reproduced in the report.

maximizing the business value of technology.[21] Chapter 14 explains how the two divide the work. The distinction rests on what each was built to do, rather than what it intends to achieve. FinOps was built to make cloud spending accountable when bills arrived after capacity had been used.[22] AI Operations Management manages consumption metered by the request and judges it against a value boundary set before spending begins.

Until Chapter 14 defines the full boundary, a provisional rule assigns activities to the two disciplines. An activity belongs to FinOps when it answers a question about the bill, and to AI Operations Management when it answers a Founding Question about the work the consumption served. Consider a constructed case in which a company's monthly AI invoice is three times the forecast. Reconciling the invoice against the contract and charging it to cost centers answers a question about the bill, so it belongs to FinOps. Tracing the increase to the workflow whose requests grew and deciding whether that workflow's output justifies its new cost answers the third and fifth Founding Questions, so it belongs to AI Operations Management. Both activities draw on the same record. Evidence that the FinOps team already traces consumption to work would move the second activity across the boundary.

Regulatory governance determines whether an organization may use AI in a particular way. This book takes that determination as given and asks how the permitted use is managed.

3.4 Five questions the discipline exists to answer

Each of the five questions corresponds to one of the discipline's functions and is answered in the chapter that teaches it. Part III covers the functions in a different order, for reasons explained there.

---

21. FinOps Foundation, "What is FinOps?," accessed October 17, 2026. The Foundation's definition, which names business value.

22. FinOps Foundation, "2024 Changes to the Definition of Cloud FinOps," accessed October 17, 2026. Cloud consumption as variable and tracked after the fact.

An organization that manages its AI consumption can answer five questions from its own records:

Question 1. “Are you paying for capability you do not need, or starving work that needs more?” This is the sourcing question, answered in Chapter 7.

Question 2. “What do you expect the AI flow to consume next period, and how will you know when it deviates?” This is the planning and budgeting question, answered in Chapter 10.

Question 3. “Who consumed what last month, in service of which work?” This is the metering and attribution question, answered across Chapters 8 and 9.

Question 4. “When capacity is constrained, who decides what runs first, and by what rule?” This is the allocation question, answered in Chapter 11.

Question 5. “Where, exactly, does this AI flow pay for itself, and who is on the hook for knowing?” This is the value boundary question, answered in Chapter 12.

These are the Founding Questions, which an organization must answer “from its own records.” It can always produce an answer by asking the provider, estimating, or repeating the business case. But the answer counts only when the organization retains enough evidence to check it, such as a provider’s figures reconciled against its own or an estimate with its assumptions written down.

At the end of Part I, readers do not yet have a method for answering these questions from records. Parts II through IV provide it: Part II presents the science the functions rest on, Part III builds the five functions, and Part IV turns them into an institution.

CRAFT SECTION

The function audit

The **function audit** identifies which of the five functions an organization performs for a deployment and which the available evidence does not show anyone performing. It builds on Part I's two earlier methods: Chapter 1's consumption-event inventory lists what a deployment consumes, and Chapter 2's three-flow mapping shows which flows are managed. The audit asks who performs each job the discipline requires, using the following six steps.

Step 1. List the deployment's AI activities: buying capacity, choosing models, forecasting spending, recording use, assigning costs, routing work, and judging the return. Start from the consumption-event inventory.

Step 2. For each activity, name the Founding Question it helps answer and assign it to that question's function.

Step 3. Use Table 3.1 to assign any activity that answers a neighboring discipline's question to that discipline.

Step 4. Mark each function as performed, partly performed, or not shown. A public record shows only what a company chose to report. If a function is missing from that record, the evidence does not show it; the omission does not prove the function absent. Only an organization's own records can establish that no one performs a function.

Step 5. For each function performed or partly performed, name who performs it. Write "not named" where the evidence identifies no one.

Step 6. State the first gap to close and the record needed to close it. Begin with the gap where the largest cost goes unmonitored, using the records the organization already holds.

Steps 2 and 4 do most of the work. Step 2 turns a list of activities into a map of the discipline, while Step 4 identifies the gaps: functions in which costs can accrue without anyone monitoring them.

Two rules keep the assignments consistent. Choosing what capacity to buy is sourcing, even when one model serves all the work. The allocation and routing function begins only when a stated rule sends different work to different sources of capacity, decides what runs first, or decides which users receive capacity and which do not.

> **DATED: JUNE TO JULY 2026**
>
> Lindy is a company of more than two dozen people that sells AI assistants for email and calendars. In 2026, Flo Crivello, who leads the company, said its largest expense was the AI models it bought from Anthropic: "Like, more than payroll." In June 2026, he announced a move to DeepSeek V4, a model from a Chinese developer, and said it would save millions of dollars.[23] Lindy's own technical account says most of its managed-agent traffic moved, while selected uses stayed on Anthropic's Sonnet model.[24]

Applied to Lindy using the public record, the procedure works as follows.

Step 1. The record shows four activities: running the assistants that customers use; choosing a model provider; knowing the total cost of the models well enough to compare it with payroll; and moving most traffic to one model while keeping selected uses on another.

Step 2. Choosing a provider answers the sourcing question. Keeping selected uses on Sonnet while the rest moves to DeepSeek sends different work to different capacity, so allocation and routing is partly performed: the sources describe the split, not the rule behind it. Knowing the total cost answers part of the metering and attribution question, but the sources do not show that cost assigned to customers or products.

---

23. "American AI is expensive. Some startups are turning to cheap Chinese models," *NPR*, July 15, 2026, accessed October 17, 2026. The cost and payroll comparison, the move, and the expected savings.

24. Bruno Škvorc, "Migrating from Claude to DeepSeek," *Lindy blog*, June 9, 2026, accessed October 17, 2026. The company's account of what moved and what stayed.

Step 3. No activity in the record belongs to a neighboring discipline. Whether a company may use a model from a particular developer or country is a question for regulatory governance, whose determination this book takes as given.

Step 4. Metering and attribution, and allocation and routing, are partly performed. Planning and budgeting and holding value to account are not shown. The cited sources contain no forecast of consumption and no measure of whether each customer's use pays for itself.

Step 5. Crivello performs sourcing in public. The sources cited here name no one responsible for the other functions.

Step 6. The first gap is attribution. A company whose largest cost is AI consumption needs to record that consumption by customer account so it can compare each customer's use with what that customer pays. The same record provides the basis for a budget and a value boundary.

Using six steps and a few public facts, the audit examined a company moving most of its AI work to a new model, a move its leader explained by cost. The public record provided no answer to two of the five Founding Questions and only part of an answer to two others. Whether the sourcing decision was sound depends on the new model's quality and the cost of the move, which the record does not show. The audit also identifies what the switch alone could not establish: which customers account for the company's new costs.

Chapter summary

AI Business Economics is the science of consuming AI: an ordered body of conditional claims, each stating when it applies, what it concludes, and which other claims it depends on.

The book presents the results of the science in numbered panels. Each panel carries an identifier and is followed by an explanation in plain business language. The full statements are in Appendix A.

AI Operations Management applies the science through five functions: sourcing, planning and budgeting, metering and attribution, allocation and routing, and holding value to account within a value boundary. It begins when a model becomes a service an organization consumes and shares boundaries with AIOps, MLOps, FinOps, and regulatory governance, while remaining a distinct discipline.

The function audit assigns a deployment's AI activities to the five functions or to a neighboring discipline and marks each function as performed, partly performed, or not shown. Its first output identifies the gap to close and the record needed to close it.

The Founding Questions are the five questions the discipline exists to answer from an organization's own records. By the end of Part I, readers understand the questions but do not yet have a method for answering them.

Key terms

AI Business Economics

The science of the business economics of consuming AI: an ordered body of conditional claims, each stating when it applies, what it concludes, and which other claims it depends on.

AI Operations Management

The management discipline that acts on AI Business Economics: the practice by which an organization buys AI capacity, plans its use, records and assigns consumption, allocates capacity, and holds that consumption to account.

Sourcing

The function that matches the AI capacity an organization buys to the requirements of its work: which models and providers serve which work, and on what terms.

Planning and budgeting

The function that states how much AI an organization expects to consume in a period and detects when actual consumption departs from that expectation.

Metering and attribution

The function that records what was consumed and assigns each unit of consumption to the team, product, or purpose it served.

Allocation and routing

The function that uses a stated rule to decide what runs first when capacity is limited and to send each piece of work to suitable capacity.

Holding value to account

The function that judges whether a deployment pays for itself within a value boundary set before spending begins and names one person responsible for knowing the answer.

Value boundary

The boundary within which an organization judges whether a deployment pays for itself, defined by the costs and value it counts, the period it covers, and the outcomes against which it measures performance.

Founding Questions

The five questions AI Operations Management exists to answer from an organization's own records, one for each of its five functions.

Function audit

The six-step method for finding which of the five functions an organization performs for a deployment: list its AI activities; assign each to a function or a neighboring discipline; mark each function as performed, partly performed, or not shown; name who performs each; and state the first gap to close.

Discussion questions and problems

DISCUSSION QUESTIONS

1 Choose one of the five dated boxes in Section 3.2. What would the organization described there have to record in its own systems to answer that function's Founding Question? Name the record and who would keep it.

2 Of the four borders in Table 3.1, which is hardest to defend, and why? State what evidence would move that border.

3 Restate one Founding Question in language your chief financial officer would use, preserving all its content. Then identify the part you were most tempted to omit.

4 Chapter 2 diagnosed a deployment through its three flows. Match each flow to the functions in Section 3.2 that act on it, then explain which flow draws on the most functions and why.

PROBLEMS

P1 · WORKED

An audit of the budget that ran out

Chapter 2's opening case described Uber's deployment of Claude Code, which consumed the company's annual AI budget by April 2026. Using only the facts reported there, apply the six steps of the function audit.

Step 1. The reported activities are buying access to Claude Code; rolling it out to roughly 5,000 engineers from December 2025; setting an annual AI budget; tracking adoption, which reached 32 percent of the engineering organization in

February and 84 percent in March; and totaling spending. Reported spending ranged from $150 to $250 a month for the average engineer and from $500 to $2,000 for the heaviest users.[25] A sixth activity was asking about value. The president and chief operating officer said the company could not yet connect its growing use of the tool to the consumer features it was producing: “That link is not there yet.”[26]

Step 2. Buying access to one tool for all engineers is sourcing, and setting the annual budget is planning and budgeting. Tracking adoption and totaling spending are part of metering and attribution. The chief operating officer’s question belongs to holding value to account, though the reporting provides no answer.

Step 3. The reporting does not place any of these activities within a neighboring discipline.

Step 4. Allocation and routing is not shown: the reporting describes no rule for which work went to the tool or any limit on its heaviest users. Metering and attribution is partly performed: spending was totaled, but the reporting does not assign it to teams.

Step 5. Engineering ran the rollout. The public account names no one who set the budget or compared spending with it before it ran out, so the audit records “not named” for both.

Step 6. The first gap is a monthly comparison of total consumption with the plan. The reported spending totals already make that comparison possible without a new record. Recording consumption by team comes next, to show where a deviation originated.

P2 · COMPLETION

An audit with three steps left open

Chapter 2’s problem P3 described the Australian Government’s whole-of-government trial of Microsoft 365 Copilot, which ran from January to June 2024. Around sixty agencies took part. Each agency nominated or approved the

25. Janakiram MSV, “Uber burns its 2026 AI budget in four months on Claude Code,” *Forbes*, May 17, 2026, accessed October 17, 2026. Rollout, adoption, and the cost per engineer.

26. “Uber burned through its entire 2026 AI budget in four months,” *Fortune*, May 26, 2026, accessed October 17, 2026. The chief operating officer’s statement.

employees who would receive licenses, and the government issued several thousand. The product was priced per seat rather than by unit of use.[27] The Digital Transformation Agency commissioned an evaluation, delivered it jointly with an external firm, and published the full report. Most participants who answered the survey after the trial reported using the tool a few times a week or less, and respondents reported saving time when they used it. These were their own assessments, and the published methodology reports no objective measure of task time or output before the trial.[28] Steps 1 to 3 are worked through below. Complete Steps 4 to 6.

Step 1. The activities are buying several thousand per-seat licenses; agencies choosing which employees receive them; commissioning an evaluation delivered jointly with an outside firm; publishing it in full; and collecting participants' reports of their use and the time it saved.

Step 2. Buying per-seat licenses is sourcing. Choosing which employees receive a license is allocation: agency nomination decided which users received capacity and which did not. Commissioning and publishing the evaluation is holding value to account within the boundary the evaluation stated, including its limits. Collecting participants' reports of their use provides a partial record of consumption and is part of metering and attribution, though it does not record each use.

Step 3. Any review of whether agencies were permitted to use the tool belongs to regulatory governance, and the audit sets it aside.

Step 4. Mark each function as performed, partly performed, or not shown. Consider how per-seat pricing changes which functions the trial needed.

Step 5. Use the facts above to name who performed each function. Write "not named" where the facts identify no owner.

Step 6. State the first gap you would close if the trial became a permanent deployment and the record needed to close it.

---

27. Microsoft, "Expanding Copilot for Microsoft 365 to businesses of all sizes," *Microsoft 365 Blog*, January 15, 2024. Per-seat pricing of the product.

28. Digital Transformation Agency and Nous Group, "Evaluation of the whole-of-government trial of Microsoft 365 Copilot," *Australian Government Digital Transformation Agency*, 2024. The trial, its evaluation, and the evaluation's stated limits.

P3 · INDEPENDENT

Whose decision is it

Consider a constructed situation. A company's FinOps lead and its new head of AI operations both claim the right to decide whether to commit to an annual volume contract with a model provider. The FinOps lead argues that the commitment is a spending decision, like reserving cloud capacity. The head of AI operations argues that it determines which work the company can afford to run. Write one page assigning each part of the decision to one discipline, using Table 3.1 and the five functions. State the evidence that would move one part across the boundary.

Interleaving: question 4 requires Chapter 2's three flows, problems P1 and P2 audit Chapter 2's cases, and the cumulative case below requires both earlier chapters.

PART I CUMULATIVE CASE

CUMULATIVE CASE · PART I

One announcement, three readings

DATED: FEBRUARY 2024. THE COMPANY'S OWN ANNOUNCEMENT, USING ITS OWN FIGURES. ONLY THAT ANNOUNCEMENT IS USED HERE.

In February 2024, the payments company Klarna announced results from the first month of an AI assistant powered by OpenAI. The company reported that the assistant had handled 2.3 million conversations in that month, accounting for two-thirds of its customer-service chats. It estimated that the assistant was doing the work of 700 full-time agents. Klarna also reported that customers resolved their questions in under two minutes during that month, compared with eleven minutes previously. It estimated that the assistant would improve its profit by $40 million in 2024.[29]

Work from this announcement alone.

First, apply Chapter 1's consumption-event inventory. Identify the consumption events the announcement implies and where each would be metered. Distinguish what the announcement reports from what you assume.

29. Klarna, "Klarna AI assistant handles two-thirds of customer service chats in its first month," February 27, 2024, accessed October 17, 2026. All figures in this case are the company's own, as announced.

Second, apply Chapter 2's three-flow mapping. Diagnose the usage, record, and cost-and-value flows, identifying the evidence behind each diagnosis.

Third, apply each of the five Founding Questions to the announcement. For each question, state whether the announcement provides an answer from the organization's own records and what record would be needed to answer it.

Keep your answers. The case returns in Chapter 6.