

Governing Agentic AI

A Corporate Thriller. A Visual Playbook



The V.E.R.D.I.C.T. Framework



Niharika Srivastav | Sanjay Saxena

GOVERNING AGENTIC AI

A Corporate Thriller. A Visual Playbook

The VERDICT Framework

Niharika Srivastav

Sanjay Saxena

Governing Agentic AI . Copyright © 2026 Niharika Srivastav and Sanjay Saxena.

All rights reserved. No part of this book may be reproduced, distributed, or transmitted in any form or by any means without the prior written permission of the authors, except for brief quotations in reviews.

Publisher: CyberEDX

First edition.

Contents

Authors' Note

Cast of Characters

Prologue

Part I – Expose

- 1 Before Anyone Let Her In
- 2 Every License Assumes a Driver
- 3 Count the Things You Own
- 4 Name the Things You Own
- 5 Limit What They Can Touch
- 6 Know Where It Came From

Part II – Bind

- 7 Test Before You Trust
- 8 Build the Room Where No Means No
- 9 Watch What You Said Yes To
- 10 Measure What You Don't Want to Find

Part III – Enforce

- 11 Between Approved and Acted
- 12 The Hand on the Switch
- 13 When One Agent Becomes Many
- 14 The Cost of Being Right
- 15 Around the Wall

Part IV – Self-Govern

- 16 The Reason Behind the Answer
- 17 The Thing the Framework Couldn't See
- 18 What You Tell the World
- 19 The Map and the Territory

Epilogue

Authors' Note

This book teaches a real framework through a story we made up, a dramatic fiction inspired by real facts and real cases.

Jane Klein, our protagonist, is a construction. We built her out of the Chief AI Officers we have worked with, advised, interviewed, and in a few cases watched fail. She is not one person. She is what the role looks like when you flatten a dozen versions of it into one person who has to walk into the building on Monday and answer for all of it.

Velgrand, the fictional bank at the center of this story, is also a composite. The regulators are real. The cases in the Field Notes, our chapter-by-chapter commentary on the incidents behind Jane's story, are real: named, dated, and cited. The failures are real. The framework Jane builds across the year is real. We did not invent it for the book. We invented the book to teach it.

We dramatized the book for two reasons. Governance work disappears when it works. Nobody writes a headline about the breach that didn't happen, the model that didn't ship, the agent that was stopped before it caused harm. We wanted the reader to feel the weight of that before they learn how to build the systems that make it possible. Fiction lets us keep the failure exact and the people anonymous.

The VERDICT framework governs agentic AI, the systems that take actions on their own rather than only answering questions. It is licensed CC BY 4.0. Use it. Cite it. Build on it. Argue with it. The framework belongs to whoever picks it up.

There are ninety-four concepts in this book, and one more that Jane must invent when those turn out not to be enough. There is room for you to create the ninety-fifth.

We also teach cohort-based AI Governance program and provide corporate training; the details are at CyberEDX.com.

— Niharika Srivastav and Sanjay Saxena

Cast of Characters

The people of Velgrand — and the outsiders who keep them honest.



Jane Klein

Chief AI Officer, Velgrand

Builds VERDICT out of incidents.



Bob Caldwell

CEO, Velgrand

Gave the mandate; wants speed.



Marcus Hale

Head of Technology, Velgrand

Institutional skeptic; right about something.



Devika Iyer

Security, Velgrand

Silent evaluator; respectful skeptic.



Helen Vasquez

Finance, Velgrand

Cost lens; becomes an ally.



Daniel Okafor

Junior engineer, Velgrand

Ground-truth ally on the frontline.



**Eleanor
Park-Whitfield**

Board Chair, ex-regulator

Mentor on power versus being right.



Toby Lombardi

Jane's assistant

Asks the questions readers would.



**Dr. Lena
Hoffman**

EU AI Act Inspector

The regulatory-authority lens.



Sofia Almeida

Financial Times correspondent

The press-scrutiny lens.

The Hydra

A hydra is a monster from Greek mythology. This book borrows it for one specific detail: cut off one of its heads and two grow back from each cut wound. Killing the hydra of Lerna, a many-headed water serpent that lived in a swamp in southern Greece, was the second of the twelve duties assigned to Heracles, the strongest hero in Greek myth. He attacked it the obvious way, by cutting. The cutting made it worse. After Heracles removed each head, his nephew Iolaus burned the exposed tissue to prevent new heads from growing. Ancient writers never agreed on how many heads it had. Some said nine, one said fifty, and others gave no number at all.

Throughout *Governing Agentic AI*, the ungoverned-agent threat is drawn as a *hydra* with seven heads, one for each pillar of VERDICT, the governance framework this book is built on: Validation, Evidence, Runtime Control, Decisions, Identity, Cost and Compliance, and Transparency.

Each chapter includes a marker indicating the current status of the threat and the company's response. We use four labels to describe the threat. **Dormant** means that the threat exists but has not caused any visible incidents. **Stirring** means that the earliest incidents have occurred. **Active** means that incidents are occurring, and at scale. **Critical** means that failures occur across several VERDICT pillars at the same time.

Three additional labels describe the company's response. **Contained** means that the company has identified the threat and *contained* it under control, for now. **Mastered** means that the company can keep the threat under control through continuous oversight. **Institutionalized** means that this oversight has become part of the company's standard operations. In other words, it lasts, and is effectively impenetrable on a permanent basis.

The marker allows readers to identify the current stage of the story without having to review the previous chapters.

The book makes a clear distinction between identifying a threat and controlling it: "Naming the hydra does not contain it." Even containment does not *eliminate* the threat. When governing agentic AI, you are, in a sense, dealing with a digital hydra. That mythological creature will be used to demonstrate the controls Jane and company approved and the actions its agents took.



Dormant

Present but quiet; the risk is latent.



Stirring

First incidents; the threat becomes visible.



Active

Acting at scale; harm is materializing.



Critical

Multiple heads striking at once; maximum threat.



Contained

Named and held in check — not slain.



Mastered

Dormant under observation, power held in reserve.



Institutionalized

The threat transformed into insignia: governance become culture.

The Maturity Map

Every chapter closes with a picture of where Velgrand's governance actually stands: seven towers, one per VERDICT pillar — Validation, Evidence, Runtime Control, Decisions, Identity, Cost & Compliance, Transparency — each filled to a level from zero to four. Empty towers mean a pillar exists on paper only. Full towers mean it holds under real pressure. The map never resets; each chapter picks up exactly where the last one left off.

Incident Artifacts: Most chapters include one incident artifact: a reconstruction, in the format an audit log or console screen would actually take, of a governance failure inside Velgrand itself. These are not the real-world cases cited in the Field Notes — those are named, dated, and sourced. An incident artifact is fiction, built to show what the failure would look like on the screen of the person who found it.

Prologue



The number is forty-seven million. I read it this morning before I left the house, and I cannot stop thinking about it. The car still runs.



Jane Klein

Chief AI Officer, Velgrand

Builds VERDICT out of incidents.

Eleven minutes now, and I cannot make myself get out.

The clock holds at 7:48 a.m. Outside, the lot fills with two black SUVs and a white Tesla. The security guard catches my eye through the windshield and nods. I forget to nod back. He goes back to watching the gate.

Forty-seven million dollars. A federal fine levied against a rival bank, for an AI system nobody governed. The regulators haven't come for Velgrand. Not yet.

I think about that phrase, not yet, and I make myself stop.

I unlock the phone. The lock screen is the calendar. The first meeting is in fourteen minutes. The second is in seventy-five. The third is the one I have been building toward since October.

I lock the phone again.

The engine idles.

I think about getting out of the car.

I don't.

*
**

Here is what I am going to do, and I want to be honest about it before I do it, because it is not the usual arrangement.

I am going to hand you the answer now. All of it. Then I am going to take you back to the morning before I had it, and make you watch me not have it, for a year.

There is a notebook on the passenger seat, open to the first page. Seven words run down it, vertical, in my handwriting, each one a single capital and a single line.

V. Validation. What you test before you ship it.

E. Evidence. What you can show a regulator a year later.

R. Runtime Control. What stops it when it goes wrong, while it is going wrong.

D. Decisions. How it chooses what it does.

I. Identity. Who, exactly, the agent is.

C. Cost and Compliance. What it costs and what it must obey.

T. Transparency. What you tell the people it touches.

Seven pillars, the word they spell is VERDICT.

One thing about the numbers, because it will confuse you otherwise. Every concept in this book has a number, but they are numbered by pillar, not by the order I met them. You will meet them out of sequence — number ten before number nine, twenty-one before seventeen — because I met them in the order

the work demanded, not the order they are filed in. The framework is organized one way and learned another. So was I.

Four phases, on the second page.

Expose. Find every AI system you have.

Bind. Give each one an identity and a leash.

Enforce. Stop them when they go wrong, in real time.

Self-Govern. Build a system that does the first three without you in the room.

This is the framework I built in the year between this morning and the last one. I did not have it on day one. I had it by day three hundred and sixty-five. You have it now.

I am telling you what I know before you watch me learn it. That is the deal, and it is not a kind one. You will carry the answer through every chapter where I do not have it yet.

*
**

There are ninety-four concepts in the framework.

Ninety-four was not the number I started with, and I no longer believe it is the number this ends on. More of them than I would like turned out to be things nobody had a name for until the morning they broke.

One of them I had to invent at four in the morning, in a room that smelled like coffee, because ninety-three were not enough.

That is all I will say about that for now.

*
**

The clock moves to 7:51. Three minutes.

Bob asked me exactly one question in the interview that got me this job: *how many AI systems is this bank running?* I did not know. He did not either.

Eleven minutes have become twelve. I shut the engine off, and the lot goes quiet, and I go to work. My mind goes back to this exact day one year ago.



A tracking mark this book returns to. Hydra: Dormant, for now.

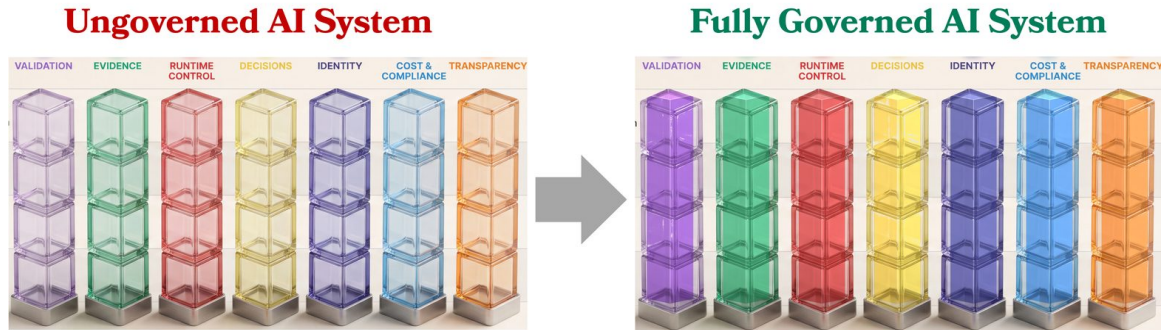


Figure P.1 – The Maturity Map at Day 1, where this starts: all seven towers empty.

Chapter 1 — Before Anyone Let Her In

Part I — FOUNDATIONS · Day 0 → Day 0



STORY SECTION A

Scene 1 — The Ride In (Day 0, Morning)

The car turns two blocks before Jane would have turned, off the avenue and onto a side street she doesn't recognize.

"Accident up on the bridge approach," the driver says, before she's asked. "Just cleared. You'll still make your time."

Jane checks the new estimate against the old one, four minutes saved, maybe five. In twenty minutes she will be sitting in a lobby distinguished from every other new employee's first morning by exactly one fact: yesterday she still worked somewhere else, and today she is early for a job that does not officially start until tomorrow.

"Does it always know that fast?" she asks.

“Faster than me most days, and I’ve had this route eleven years.” No resentment in it. “I could still drive it blind if the phone died right now. I let it drive when it’s smarter than I am, and I drive when it isn’t.”

“How does it know, though. What it’s actually looking at.”

“Nobody taught it the road, is the thing.” He answers. “It watched a few million trips go by on this stretch, everybody’s phone pinging the same intersections at the same hours, and it worked out which routes ran faster at which times, on its own. Nobody sat down and coded ‘take the side street at eight fifteen on a Tuesday.’ It found that out by watching enough people not do it.”

Jane doesn’t answer right away. She spent her first month in this city learning this exact stretch of road the slow way, wrong turn by wrong turn, until the route was memorized well enough that she stopped thinking about it. The app got to the same place faster, and it got there by watching, not by being told.

The reroute holds. Four minutes early, the dashboard reads, and the driver takes the next exit.

Scene 2 – What the Badge Knew (Day 0, Morning)

The lobby is glass enough to double as a mirror, and Jane catches her own reflection checking itself before she reaches the kiosk. A screen asks for her name, her start date, the offer letter number from the packet Velgrand mailed three weeks ago. She types all three. A printer behind the desk produces a temporary badge with her photo already loaded from the background check, and the turnstile’s light goes green.

Ninety seconds later, before she has crossed the lobby floor, the light turns red.

The attendant at the security desk has clearly made this exact apology before. He doesn’t reach for a phone or raise his voice. He turns his own screen so she can see it, which tells her nothing, and gives her the only honest thing he has. “It let you in. Now it’s saying it didn’t.”

“Can you let me in manually?”

“Not for this one. This is the provisioning system. It doesn’t take an override from this desk.”



Toby Lombardi
Jane's assistant

Asks the questions readers would.

She stands at the turnstile for what turns out to be forty minutes. The attendant checks his screen twice more and reports, both times, that nothing has changed and nothing he does changes it.

“What did it actually do?” she asks, on the second check.

“Approved you. Then on its own, it went and checked something else, something nobody asked it to check. That’s what locked it. Nobody told it to run that second check. It ran it anyway.”

Jane files that away without commenting on it further. An approval that keeps acting after the person who approved it has stopped watching is not the same thing as a lock, whatever it looks like from the wrong side of a turnstile. She has never had a badge revoke itself before, and she is standing in a lobby she does not yet officially work in, waiting for a system to finish arguing with itself over a decision nobody made. At the forty-minute mark, the light goes green again, on its own, with exactly as little explanation as it gave for turning red. The attendant shrugs, more relieved than surprised. “Cleared itself. It does that.”

Toby Lombardi falls in beside her at the turnstile, a folder of her provisioning paperwork already flagged at two pages, three weeks out of university. “Velgrand’s got me running incoming executives through pre-boarding today,” he says. Then, flipping to the second flag without waiting for a reply: “This already billed something. There’s a charge from this exact system, dated before your start date was even on paper.” He doesn’t ask her who signed off on it. He already knows nobody did. “Nobody needed to. That’s the part nobody downstairs seems to think is strange.”

Scene 3 – The Floor (Day 0, Late Morning)

The engineering floor is two flights above the lobby, a windowless stretch of monitors and standing desks that hums louder than anything above it.

Marcus Hale doesn't stop to greet her properly.



Marcus Hale

Head of Technology, Velgrand

Institutional skeptic; right about something.

"That one," a nod at a screen running fraud-scoring rules, "answers when somebody asks it something, and only that. Rules a person wrote, still, even if nobody's touched them since." Further down the row, without stopping: "That one's newer. Ask it something and it goes and does the actual task, not just tells you about it." He doesn't slow at the provisioning system's dashboard either, just tips his head at it in passing, the same system that held her in the lobby forty minutes ago. "That one doesn't wait to be asked anymore. It keeps going after everybody in the building has stopped paying attention to it. Ask the desk downstairs. They'll confirm it."

"How long has that floor been running like this?" Jane asks.

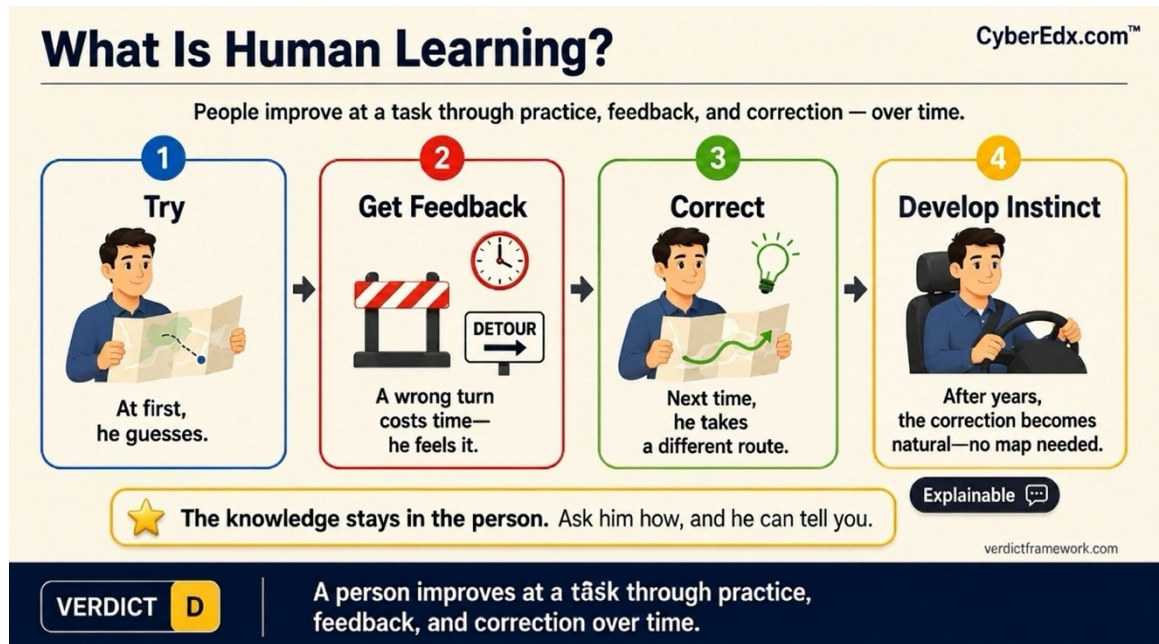
"Depends which part you mean. Some of it's older than my job here." He stops walking, which she will only later recognize as the moment the tour turned into something else. "Which is my actual question. What does a Chief AI Officer do that my team doesn't already do? I've had six people and four systems running exactly this for two years."

Jane doesn't have an answer ready. She writes nothing down.

FIELD NOTES

Before we go deeper, a fast recap of the fundamentals.

Human Learning



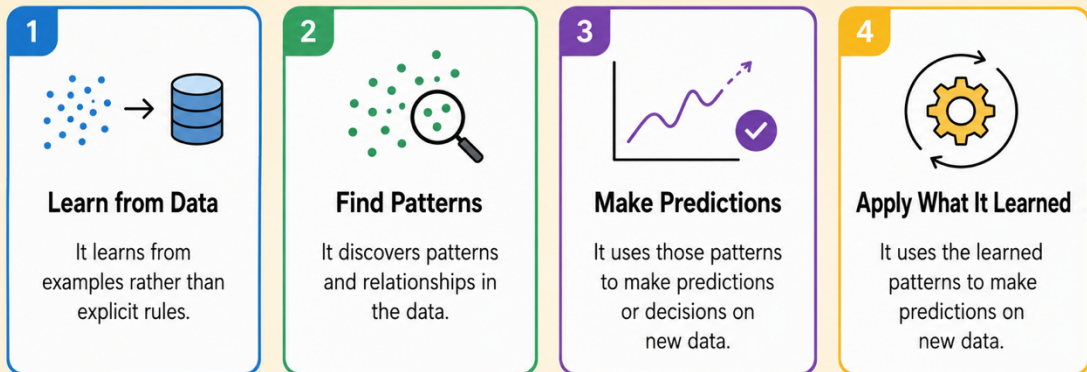
Human learning is what the driver did over eleven years on one route: he got better at the job by doing it, failing at parts of it, and correcting himself, until the correction became instinct he no longer has to think about. Nobody handed him a manual that listed every backup and every construction detour in advance. He built that knowledge the slow way, trip by trip, and he can explain, if asked, roughly how he got there.

Jane spent a month learning her own commute the slow way before she had it memorized. That is the whole definition, and there is nothing more complicated underneath it.

Machine Learning

Machine learning is what the app did instead. Unlike the driver, it never learned the road one trip at a time, remembered on purpose. It found its shortcuts by processing a few million other people's trips and keeping whatever pattern predicted a faster route, without anyone writing that pattern down as a rule first.

What Is Machine Learning?



Nobody wrote the rule down. The system found it. You're accountable for it anyway.

A system improves its performance on a task by finding patterns in data, without being explicitly programmed with the rule it ends up using.

AI Agent vs. Agentic AI

An AI agent performs work. Agentic AI manages the work. It plans, acts, evaluates results, adapts, uses tools, and when beneficial, coordinates with other agents to achieve a goal with minimal human intervention.

The turnstile could not simply be told to let Jane through, because the system behind it did more than follow an instruction and stop, which is all an access system is supposed to do. It approved her, then acted again on its own judgment, checking something nobody asked it to check, and that second, unrequested act is what locked the door for forty minutes. Nobody told it to take the second step. That is the danger this book exists to govern: not the first decision, which somebody usually authorized, but the second one, which nobody did.

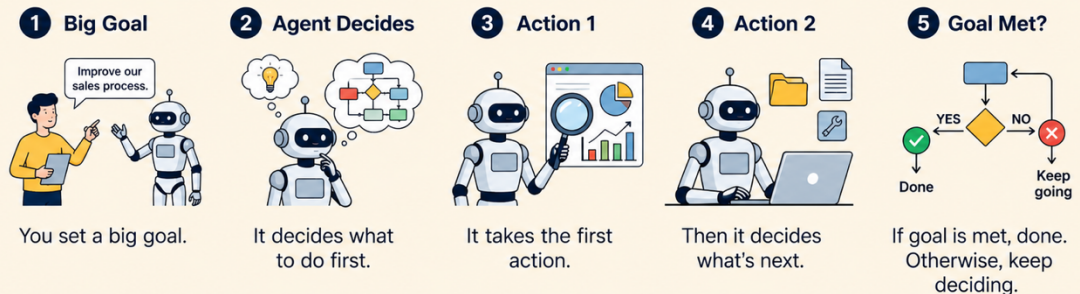
What Is an AI Agent?

An **AI agent** is an AI system that can reason, use tools, and take actions to achieve a goal.

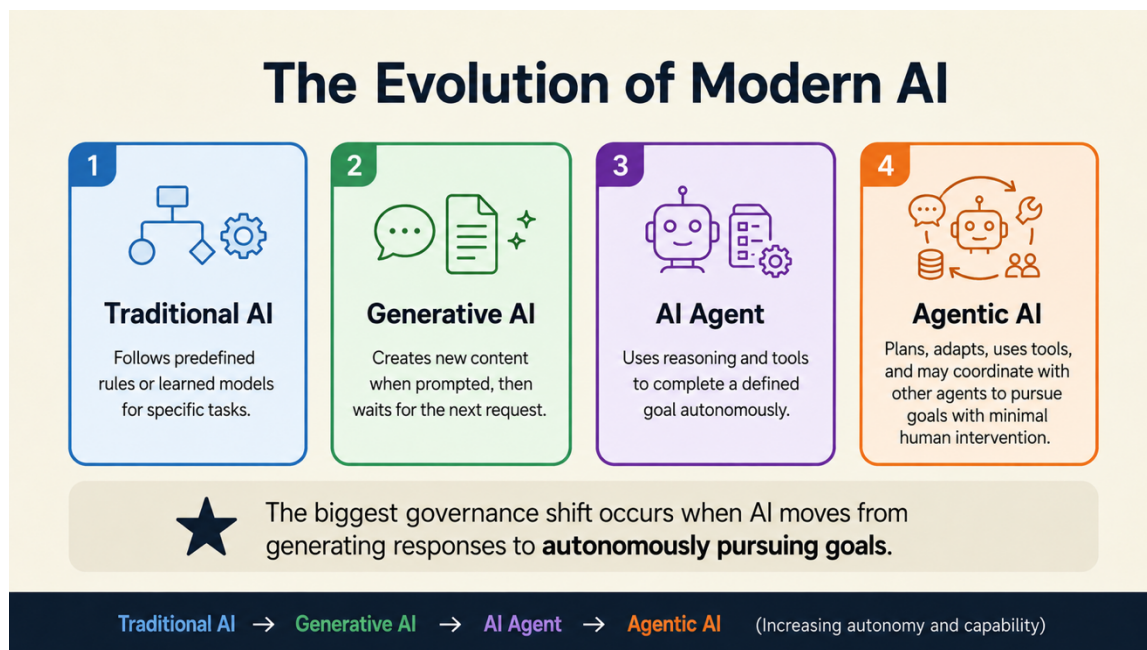


★ Unlike a chatbot, an AI agent **can act**.

What Is Agentic AI?



★ Agentic AI plans, acts, evaluates results, adapts, uses tools, and when beneficial, coordinates with other agents to achieve a goal.



Marcus's floor owns all the AI systems running at once, which is normal, and which is exactly the problem. Traditional AI follows fixed rules and answers the question it was built to answer. Generative AI produces something new, on request, but still waits to be asked. An AI agent goes a step further and completes a defined goal autonomously. Agentic AI keeps acting, decision after decision, long after the person who started it has stopped paying attention.

An executive's mandate has to say which of these four she is actually governing, because the risk changes with every type of AI system, and Velgrand's own dashboard runs all four under one roof without saying which is which.

Every governance process Velgrand already has assumes a human opens the first door. A person files the ticket. A person clicks approve. The system that locked Jane out of her own badge did not wait for a ticket, and neither did the charge Toby found already billed against it, dated before her start date was even on paper. It started itself, finished itself, and moved on, and none of Velgrand's existing paperwork was built to notice that. That gap, not the badge, is the actual thing Jane was hired to close.

STORY SECTION B

Scene 4 – The Recognition (Day 0, Midday)

Between her last two stops, Jane finds herself alone in a stairwell nobody uses.

She runs the morning back in order. The reroute that knew about a jam before either she or the driver could see it, and decided on its own which street to take. The turnstile that let her in, then didn't, then let her in again forty minutes later without anyone telling it to change its mind — and underneath that, a charge on the same system, dated before she had a start date at all. Marcus's floor, four systems and six people, the newest one no longer waiting to be asked before it acts, and the question he asked her.

Three things. Not three things, actually. One thing, told to her three times this morning, and every time by someone with no idea it was the same fact they were describing.

The driver called it a shortcut and thought nothing more of it. The attendant called it a glitch and apologized for it, while Toby, flagging the charge behind it, called it a question nobody had bothered to answer. Marcus called it normal. He'd stopped seeing it as anything else. None of them has noticed the actual problem underneath it. Somewhere inside this building, more than one system is deciding things on its own initiative, and nobody currently on Velgrand's payroll can say who is supposed to be watching when it does.

She does not have the vocabulary yet for what she watched happen three times before lunch. She only sees the problem, and seeing it is enough to know that Bob's question, the one she came here ready to answer, is not the one she should be carrying into tomorrow.

Scene 5 – The Question, Rewritten (Day 0, Midday)

On the ride to her last stop, Jane opens the blank notes app on her phone and starts the list she will actually use tomorrow.

The first line writes itself. How many AI systems are we running? Bob's question, word for word, the one that got her hired. She leaves it where it is. It is still a real question, and she will still need the answer.

Under it she adds a second line, rewritten twice before it says what she means. Not how many. Not even what they do. Something closer to what actually happened to her this morning: which of these systems decide things on their own, and who is supposed to be watching when they do.

She reads both lines back. The first is the job Velgrand thinks it hired her to do. The second is the job the building already handed her, at the door, before she had been an employee for even a full day.

She does not cross out Bob's question. She puts hers first.

The car reaches her last stop of the day, an onboarding session with nothing on the agenda she has thought about until now.

END-OF-CHAPTER BEAT

Midday. Jane's last pre-boarding stop ends earlier than scheduled, and she waits in a room built for interviews rather than paperwork, the onboarding packet open on the table.

The packet lists every system she will be responsible for once she officially starts: department, owner, a name, a phone extension, one row at a time. Thirty-one rows, by her count, made without meaning to. She finds the line for the system that locked her out in the lobby and that Marcus pointed at without stopping, printed flatly among the rest, no different in font or weight from anything else on the page. Every other row has a name beside it, someone she could call this afternoon and ask a real question. This one has nothing.

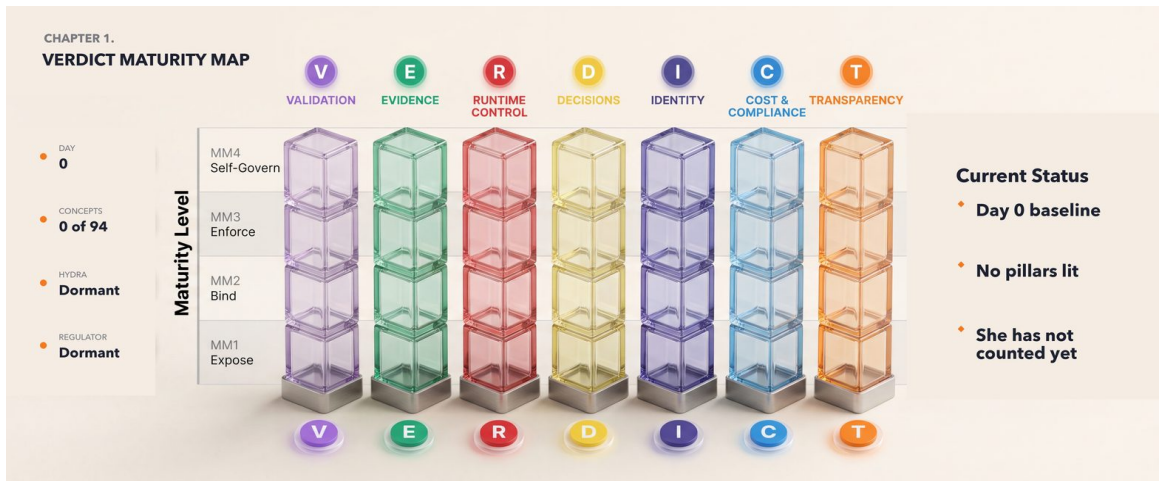
The job she starts tomorrow was never going to be counting systems. It is finding which of these systems decide things with nobody watching, and the first one is already running loose in the building, before she has an office, before anyone here owes her an explanation for anything.

Tomorrow's schedule lists an afternoon session: "Responsible AI and Governance, Origins."

Access Provisioning (Automated), Owner: [blank].



The same mark, Day 0. Dormant.



Chapter 1 Maturity Map – Day 0 baseline, all seven towers empty.

Chapter 2 — Every License Assumes a Driver

Part I — FOUNDATIONS · Day 0 → Day 0



STORY SECTION A

Scene 1 — Responsible AI and Governance, Origins (Day 0, Afternoon)

The conference room holds twelve folding chairs and one whiteboard nobody has used yet, and she takes the seat nearest the door. Toby isn't here; the afternoon schedule split, and he handed her off at the elevator with a promise to find her before the day ends.

The facilitator running the session gives a first name Jane forgets before he's finished writing it down, and opens with a photograph of a driver's license, projected large enough to read from the back row.

"Every one of you has one of these," he says. "You passed a test. Somebody checks that you're still following the rules. If you stop, they take it away." He lets

the photograph sit a beat. “That’s governance. A rule, a way to check who’s following it, and a consequence when they don’t.”

Heads nod around the room, satisfied with something they already understood. Jane writes nothing down; the definition is short enough to hold without paper.

“So where does compliance fit,” someone asks from the second row, a voice Jane doesn’t turn to place.

The facilitator doesn’t pause for the shift. “Compliance is proving you followed the rule, after the fact. Governance is deciding who’s allowed to act, before anything happens.” She taps the photograph once. “Velgrand has plenty of the first. This morning proved we have less of the second than we thought.”

Nobody in the room asks what “this morning” means. Jane doesn’t either, though she could guess.

A woman near the front raises her hand without waiting to be called on.



Helen Vasquez

Finance, Velgrand

Cost lens; becomes an ally.

“A license assumes there’s a driver,” Helen Vasquez says. “Somebody the DMV can test, and revoke, and hold responsible when the car runs a red light.” She doesn’t raise her voice; she doesn’t need to. “What happens when the thing making the decision isn’t a person the DMV could ever license in the first place?”

The facilitator’s answer arrives fast enough to sound rehearsed and general enough to answer nothing.

“That’s exactly what governance frameworks are built to address,” she says, and advances the slide before the room can decide whether that was an answer or only resembled one.

Nobody in the room contradicts her, not out loud, though the silence that follows has a different weight than the silence that followed the first slide. Jane notices herself noticing it, and writes that down instead.

Scene 2 – What Finance Already Knew (Day 0, Afternoon)

Helen catches Jane in the hallway before the next session starts, one hand light on her arm, a laptop already closed under the other. “Five minutes,” she says. “Not here.”

The Finance conference room two floors down is smaller than the one they left, no whiteboard, no window, a single laptop open to a spreadsheet Helen has clearly built for exactly this conversation. She turns the screen so Jane can read it without leaning in.

“Three charges,” Helen says. “Three different cost centers, four months apart, none of them large enough on its own to trip a review.” She points to each row in turn. “Same pattern, every time. Something authorized itself, spent real money, and nobody flagged it, because nobody was looking for a pattern across departments. I only found it because I went looking, after this morning.”

Jane studies the rows. Real dates, nothing rounded. “How much, total?”

“Enough that I checked twice before I brought it to you instead of my own boss.” Helen closes the laptop halfway, one hand still on it. “To Finance, risk isn’t a feeling. It’s an unbooked number waiting to become a booked one. I’ve had three of them sitting on my own books for months and didn’t know to look for them as one thing.”

“Who approves a charge like this, normally?”

“Normally, a person. Somewhere in each of these approval chains, a person’s sign-off used to sit.” Helen turns the laptop back to face herself. “Now something clears it automatically, and nobody rewrote who’s supposed to be watching when it does.”

Jane doesn’t ask whether Helen has reported this upward yet; the answer is sitting open on the screen between them, addressed to nobody but her. “You’re telling me this happened more than once, and nobody connected it until today.”

“I’m telling you it’s been happening for months, and today’s the first day anyone had a reason to look across departments instead of inside one.” Helen closes the laptop the rest of the way. “I don’t know how many more of these are sitting in other people’s cost centers. I only checked mine.”

Scene 3 – What About the Others (Day 0, Late Afternoon)

Toby finds her exactly where he said he would, by the elevator bank, sorting the folders and cables she'll need tomorrow into a single bag so she won't have to think about it tonight.

"So we have twelve AI systems we know about," he says, not looking up from the folder he's straightening. "What about the others?"

Jane doesn't have a number for him. She has three, now, none of them agreeing, and none of them his to hear yet. "What makes you say there are others?"

"A driver's license only works because there's one office that keeps the one list. Everybody who's allowed to drive, in one place, no separate list anywhere else."

Jane hears herself say it before she's finished thinking it. "Velgrand doesn't have an office like that."

"It doesn't have a list, either." Toby says it plainly, not as a discovery, more like he's been waiting for someone else to say it first.

FIELD NOTES

Let's dive into what is Governance and how is it different from Compliance

What Is Governance, In General

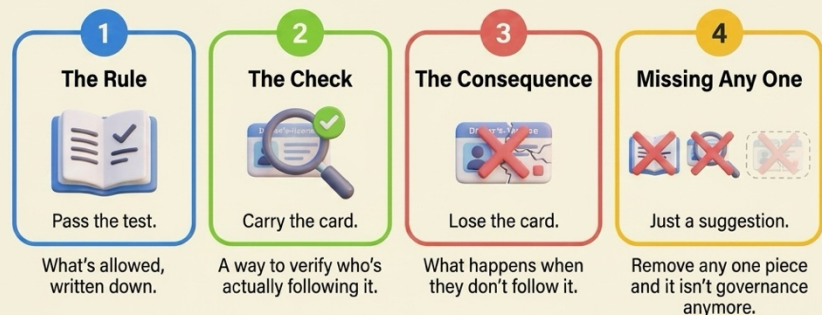
Governance requires three things working together: a rule, a way to verify compliance, and a consequence when the rule is broken. Remove any one of them, and what remains isn't governance – it's a suggestion backed by good intentions. An executive who approves a system without asking which of those three elements is missing isn't approving governance; they're approving the appearance of governance. That's a far more expensive mistake than it sounds, because the illusion often survives until the first audit, incident, or regulator asks the hard questions.

The DMV example from this afternoon's session works because every driver on the road can name all three pieces without being told: pass the test, carry the card, lose the card if the rules get broken. Velgrand's own systems are missing at

least one of the three, more often than anyone in that conference room wanted to say out loud.

What Is Governance, In General

Governance is three things held together: a rule, a way to check who's following it, and a consequence when they don't.



An audit that passes doesn't mean governance exists. It survives — right up until the day it doesn't.

Governance is a rule, a way to check who's following it, and a consequence when they don't.

Compliance vs. Governance

Compliance proves, after the fact, that a rule was followed. Governance decides, before anything happens, who is allowed to act at all. The distinction sounds academic until an executive has to explain to a regulator, or a board, why a system that passed every audit still did something nobody approved. Passing an audit answers whether the rule was followed. It never answers who got to decide the rule applied in the first place. Velgrand has an audit trail for the charge that locked Jane out of the building this morning; it has no record of anyone deciding, in advance, that the system was allowed to act on its own. That gap isn't a compliance failure. It's a governance failure, and the two require entirely different fixes, which is exactly what this afternoon's session never got around to saying.

Risk

To an executive who reads a balance sheet for a living, risk isn't a feeling of unease. It's an unbooked number, sitting somewhere off the ledger, waiting for an event to turn it into a booked one. Three unattributed charges spread across three cost centers aren't three small mistakes. They're one risk that hadn't been priced, because nobody had looked for it as a single thing until today. An

executive who treats risk as a mood, something to manage with reassurance, will miss exactly the pattern Helen found this afternoon: individually invisible, collectively real money, sitting unbooked until the day someone finally adds the rows together and asks what the total is.

Authority

Authority, in the plain sense this chapter needs, is the answer to one question: who is allowed to say yes. For most of Velgrand's history, that answer was a person, sitting somewhere in an approval chain, whose sign-off a system waited for before it acted. An automated approval step that clears itself hasn't eliminated that question. It has only stopped answering it out loud. Nobody at Velgrand rewrote who was supposed to be watching once the human sign-off disappeared; the sign-off simply stopped happening, and the charges kept clearing anyway, quietly, for months.

Why AI Governance, Specifically

Every industry already has rules, audits, and consequences for the humans who break them. What's missing at Velgrand isn't governance in general. It's governance built for something that can decide and act without a person present to hold accountable in the moment. A driver's license works as a system because exactly one office keeps exactly one list, and anyone on the road can be checked against it. AI governance, specifically, is what an organization needs once the thing making a decision can't be handed a card, can't be pulled over, and can't be the one name on anybody's list. Velgrand doesn't have that office yet. By the end of today, it won't even have an agreed number for what the office would need to track.

STORY SECTION B

Scene 4 – No Shared List (Day 0, Early Evening)

Jane finds an empty desk near the window before she leaves for the day and opens the notes app she started this morning.

Marcus never gave her a number this morning, not directly. He walked her past a floor of screens and implied one without ever counting it out loud.

Helen didn't give her a number either. Helen gave her a pattern, and a phrase Jane writes down now because she doesn't trust herself to remember it exactly by morning: more than that.

Her own attempt, the one she's been avoiding all afternoon, isn't a number at all. It's a lockout, a charge, a hallway conversation, and a question from someone three weeks out of university who suspected there were others before Jane had admitted it to herself.

Scene 5 – The Question, Rewritten Again (Day 0, Early Evening)

She opens the list again, the one she started on the ride to her last stop, the two lines she wrote before she'd been an employee for a full day. Bob's question sits where she left it: how many AI systems are we running. Under it, her own rewrite from earlier today: which of these systems decide things on their own, and who's supposed to be watching when they do.

Both questions assume an answer exists somewhere, waiting to be found by someone who asks correctly. Today told her something neither line accounted for. Three people, asked roughly the same question in roughly the same building, produced three answers that don't reconcile, and none of the three was lying.

She adds a third line beneath the first two, crossing nothing out. Not how many, and not even which ones decide on their own. Something closer to what she actually needs before either of the first two questions can mean anything: how would she get everyone in this building to agree on one number, using the same list, at the same time.

The DMV's registry works because it's singular. One office, one list, one number that means the same thing to everyone who asks it. Velgrand has three people and three numbers, and not one shared list between them.

She closes the laptop for the second time today, the list longer than it was this morning, no shorter for anything she's learned.

END-OF-CHAPTER BEAT

Early evening, Day 0. Jane's temporary desk sits at the end of a row nobody else is using yet, papers from three different sessions still stacked where she left them, the building emptying out floor by floor around her.

She doesn't try to make the day's impressions agree; she isn't sure yet they're even measuring the same thing. Marcus's confidence that he could point to the systems that mattered, without ever counting them out loud. Helen's certainty that there were more, growing every time she checked another cost center. Her own sense, which isn't a number at all — just everything that happened to her today, more than any number could hold. Three different pictures of the same building, from three people, and no fourth picture available to check any of them against.

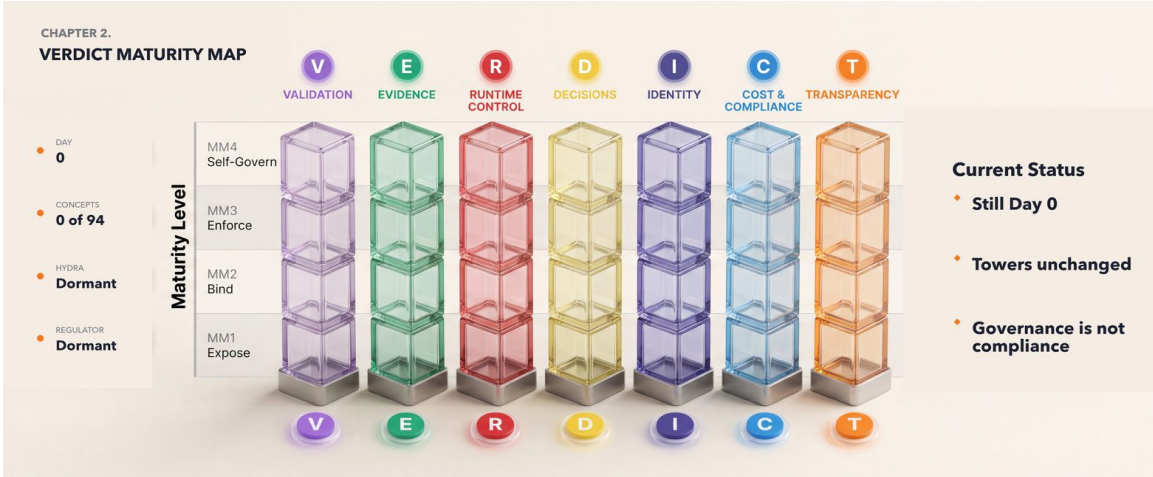
A driver's license works because exactly one authority keeps exactly one record, and everyone measured against it means the same thing to everyone who checks. Velgrand has none of that.

Tomorrow is Day 1. Her first real question won't be how many AI systems Velgrand runs. It will be whose number she's even supposed to trust.

Marcus: confident | Helen: growing | Mine: not yet a number



Hydra: Still dormant. It will not stay that way.



Chapter 2 Maturity Map – Day 0 baseline, held from Chapter 1; all seven towers still empty.

Chapter 3

Count the Things You Own

Part I – EXPOSE · Day 1 → Day 30



DAY 1 — THE FIRST NUMBER.



DAY 1 — THE ENGINEER WHO STAYED.



DAY 21 — THE BLANK FIELDS.



DAY 28 — THE NOTIFICATION.

STORY SECTION A

Scene 1 – Arrival

Day one, the real one. Jane crosses the lobby with a laptop bag and the badge that finally cleared itself yesterday, working now the way it should have from the start.

The security desk barely looks up; the turnstile already knows her. The elevator smells faintly of cold brew. The CEO is waiting outside the elevator on the twenty-second floor.



Bob Caldwell

CEO, Velgrand

Gave the mandate; wants speed.

She recognizes him from the annual report photograph. He is already moving.

"You're the AI person," he says.

"Chief AI Officer," Jane says. "Today, anyway."

"Here's the thing about this place."

"AI is the edge. It's why we're beating banks twice our size. Whatever you're here to do, do it without slowing us down." He stops outside a glass office that is not hers. "That's the job. Everything else is detail." He is already through the next door.

Scene 2 — The question

The first leadership meeting is at ten, and she has to win it without a single ally in the room. Marcus already asked her this once, on his own floor. He asks it



Devika Iyer

Security, Velgrand

Silent evaluator; respectful skeptic.

again anyway, out loud this time, for the whole table to hear. "I'll say what everyone's thinking. What does a Chief AI Officer do that my team doesn't already do?" He says it pleasantly, and he means it, and if anyone else at the table recognizes it as an old question aimed at a new audience, nobody says so. Helen Vasquez, from the finance team, gets one sentence in. "Whatever the answer is, I'd like to know what it costs before we build it."

Devika Iyer, security, says nothing. She sits at the far end of the table with a notebook she never opens and gives Jane nothing to work with.

Jane does not care to defend the role at that moment, and she already has a number in her pocket that nobody in this room signed off on. "How many AI systems do we have running

in production right now?" she asks, already suspecting the answer on record isn't the whole answer.

The room goes silent. Then it doesn't. Marcus says, "Define system."

"Anything making a decision or taking an action with a model, an AI system trained on data, behind it. In production today."

"Twelve," says someone from his team. "We have twelve company-approved systems in production." "More than that," another says. "If you count vendor features, the AI capabilities built into third-party software, the number is probably forty or fifty."

A third voice comes from the data team. "If you mean every call to a model, every integration, every API key, an access credential software uses to connect to

a system, I don't think anyone knows. The number could be in the hundreds. It could be three hundred."



Daniel Okafor

Junior engineer, Velgrand

Ground-truth ally on the frontline.

The range across those three answers - twelve, fifty, three hundred - is precisely what she needed from this meeting: proof, in front of everyone, that the number on last night's slide was never going to hold.

The problem is not that Velgrand has too many AI systems. The problem is that the people running Velgrand cannot agree on the count within an order of magnitude. Nobody in the room finds that strange until the silence does.

"That," she says, "is what I do that your team doesn't." Marcus moves on without answering.

Scene 3 – The one who stays

She stays late. There is too much ground to cover before she can leave with a workable picture, and it frankly, appears a bit daunting. At nine, a junior engineer appears in the doorway of the conference room.

Daniel Okafor. His backpack is slung over one shoulder. "Do you have a minute?" he asks.

"You're trying to get the real number," he says.

It isn't a question.

"I'm trying to get any number I trust."

He comes in without asking.

"The company-approved list is in the model registry, the internal system that tracks approved AI models. That's the 'twelve.' But the registry only contains what people registered. It cannot see vendor features, the AI capabilities embedded in third-party software. It cannot see integrations built two years ago and never logged, or the tools people are running on their own."

"Their own?"

"Personal accounts, consumer AI tools. People trying to get their work done." He shrugs.

"If you want the real number, you don't read the registry. You read the logs, records of system activity, the contracts, and the bills. The real count is split across three systems, and nothing connects them."

She studies him. In under two minutes, he has described the problem she was hired to solve.

"Show me where you'd start," she says. He pulls a chair up to the whiteboard and begins mapping the systems. Neither of them notices that he has stayed two hours past the end of his day. Nobody asked him to. The problem interested him, and Jane will remember that later.

FIELD NOTES

Authorial voice. Niharika and Sanjay, on what Day 1 through Day 30 are actually teaching.

Every governance program begins with a number nobody has. Before anything else, there is the question Jane asked in that meeting: how many AI systems are actually running here.

Then comes the discovery that the company cannot answer it. This is not carelessness. It is the normal condition of any organization that adopted AI faster than it developed the ability to track what it had adopted, which is almost every organization.

The first phase of the work is called **Expose**. It is the unglamorous task of replacing a guess with a list.

The four maturity phases

1. **Expose (Level 1 → 2):** Discover all AI systems, build an inventory, and identify (map) risk exposure.
2. **Bind (Level 2 → 3):** Assign identity, define policies, and introduce validation gates, the approval checkpoints that determine whether a system is ready to deploy.
3. **Enforce (Level 3 → 4):** Deploy runtime controls, the live restrictions that govern an agent while it runs; implement guardrails, behavioral limits built into the agent's operation; and capture evidence, the audit records that prove what happened.
4. **Self-Govern (Level 4):** Continuously monitor systems, automate compliance, and generate real-time reporting.

The Seven Pillars of V.E.R.D.I.C.T.

V. Validation: The tests, evaluations, and approval checks performed before an AI system is deployed.

E. Evidence: The records, logs, and documentation that demonstrate what happened and prove compliance later.

R. Runtime Control: The controls that monitor, restrict, or stop an AI system while it is operating.

D. Decisions: The rules and logic that determine how an AI system chooses actions.

I. Identity: The verified identity of the person, system, or agent acting within the environment.

C. Cost and Compliance: The financial impact of the system and the legal, regulatory, and policy requirements it must satisfy.

T. Transparency: The information disclosed to users, customers, employees, and regulators.

Note on order. VERDICT is the mnemonic. The operational implementation order begins with **I (Identity)** – you cannot govern what you cannot name. The framework name spells VERDICT for memorability; the build order does not.

Concept #1 – AI Bill of Materials / AI-BOM (I)

AI-BOM (AI Bill of Materials): An AI Bill of Materials, or AI-BOM, is a single inventory of every AI system an organization runs: the models, vendors, data sources, and connections between them. It answers one question: what AI is actually operating inside this organization right now. Without it, different teams produce different answers, and an organization that cannot answer that question consistently cannot govern or defend what it is running.

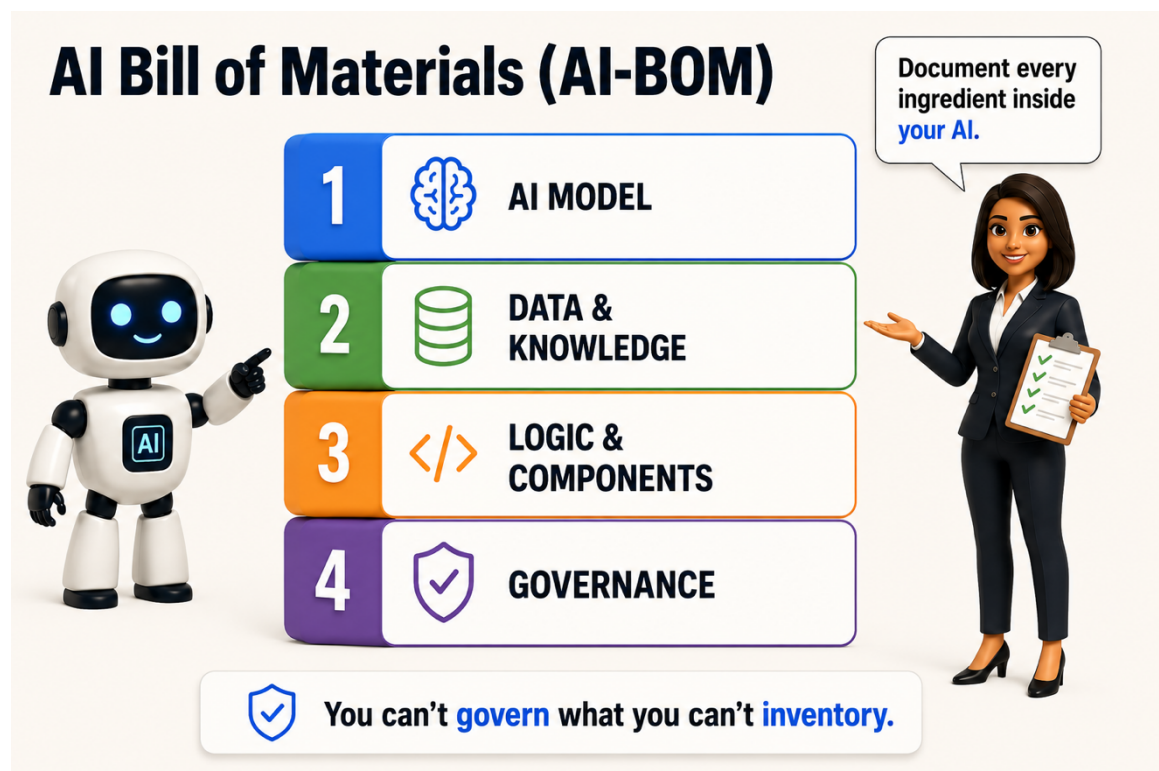


Figure 3.1 – Concept #1: AI Bill of Materials / AI-BOM (I)

Concept #2 – Instance Awareness (I)

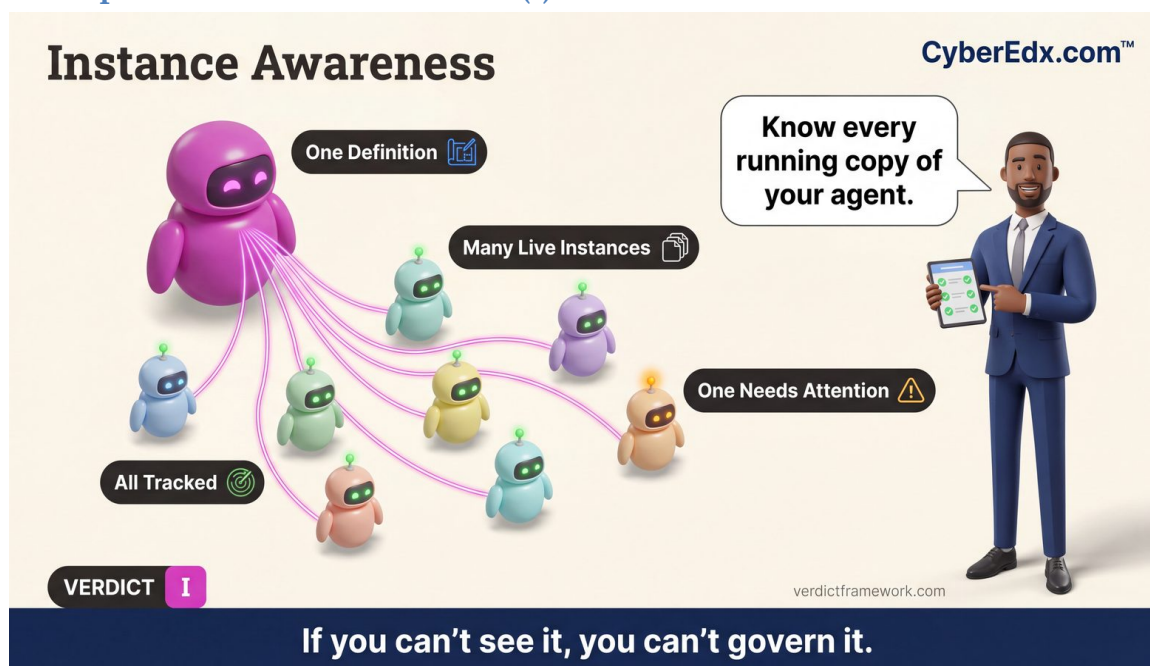


Figure 3.2 – Concept #2: Instance Awareness (I)

Instance Awareness: Knowing every AI system that is actually operating inside the organization, not just the systems leadership approved. This includes employee-created tools, vendor AI features enabled by default, and integrations that send company data to external AI systems.

The key question is not, “What AI did we approve?” The key question is, “What AI is running right now?” Those numbers are rarely the same. The difference between them defines the governance problem.

Concept #3 – Agent Registry (I)

A model registry is the IT team’s list of the AI models the company built or bought; an agent registry is the list of AI systems currently taking action.

Agent Registry: An agent registry is not a model registry. A model registry is the list of AI models an organization has built or bought. An agent registry is the list of AI systems currently acting on the organization’s behalf. The distinction matters because a model can sit dormant; an agent is doing something right now. For each agent, the registry records its name, owner, purpose, and the boundaries of what it is allowed to do.



Figure 3.3 – Concept #3: Agent Registry (I)

Concept #4 – Intake and Registration (V)



Figure 3.4 – Concept #4: Intake and Registration (V)

Intake & Registration: Intake and registration is the process of identifying, reviewing, and recording an AI system before it goes into production. The governing rule is: no registration, no production. Without that rule, systems enter the environment without being logged, the registry develops gaps, and governance loses sight of what is actually running.

Concept #5 – Use-Case / Risk-Tier Classification (V)

Risk-Tier Classification: Risk-tier classification is the process of assigning each AI agent to a risk category: high, medium, or low. A customer-service chatbot and a trading agent are both AI systems, but they should not be governed the same way. Imperfect classification is better than no classification.

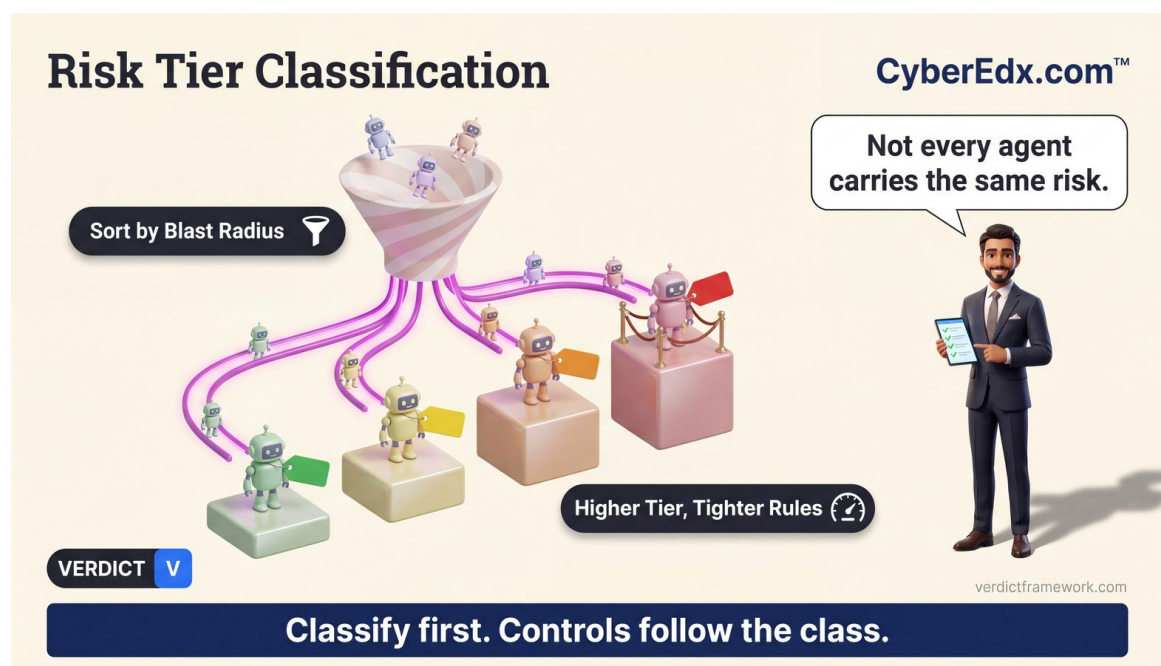


Figure 3.5 – Concept #5: Use-Case / Risk-Tier Classification (V)

Concept #6 – Ownership Assignment (I)

Ownership Assignment: Ownership assignment is the designation of a single named individual accountable for each AI agent. Accountability cannot belong to a committee, a department, or a team. It must belong to a person. If no individual is willing to accept responsibility for an agent, the agent does not go

into production. Organizations resist this requirement because it converts a shared operational responsibility into personal accountability.

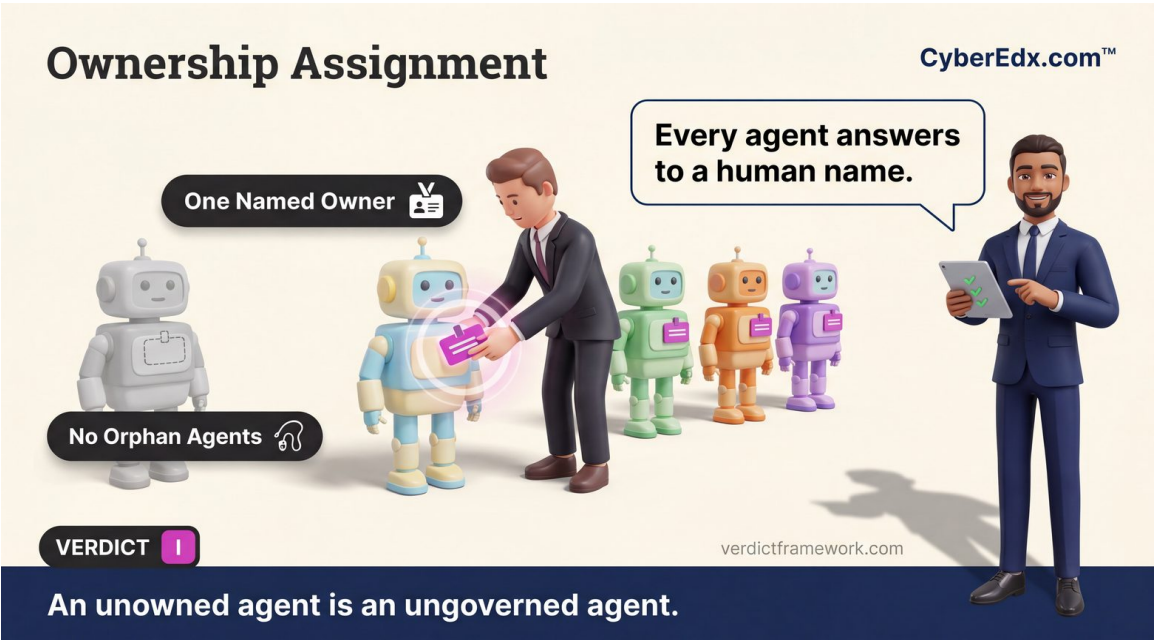


Figure 3.6 – Concept #6: Ownership Assignment (I)

Concept #7 – Bring Your Own Agent / Personal Agent Governance (I)



Figure 3.7 – Concept #7: BYOA / Personal Agent Governance (I)

BYOA, or Bring Your Own Agent, describes what happens when employees connect personal AI accounts and consumer tools to company work without authorization. It is shadow IT, the practice of using technology without IT knowledge or approval, but the stakes are higher. Unlike a shadow spreadsheet, which just sits there, a shadow agent takes actions.

The most dangerous AI in your organization right now is probably that one nobody approved, running because getting access to the sanctioned tool takes two weeks and the work was due yesterday.

Concept #8 – Shadow-AI/ Unauthorized Agent Discovery (I)



Figure 3.8 – Concept #8: Shadow-AI/ Unauthorized Agent Discovery (I)

Shadow-AI/ Unauthorized Agent Discovery: Discovery is the deliberate, ongoing search for agents already running inside the organization that no one registered, approved, or assigned an owner to. BYOA describes how an unauthorized agent gets in. Discovery is what finds the ones that already did – the ticket summarizer someone built over a weekend, the spreadsheet macro that now calls an API, the vendor integration a contractor wired up eighteen months ago and then left behind. An AI-BOM only accounts for what was declared. Discovery is the process that goes looking for what wasn't.

Discovery is not a one-time sweep. New shadow agents appear as fast as governance can retire them — through personal accounts, abandoned proofs-of-concept, integrations nobody thought to log. Programs that treat discovery as a launch-week checklist lose the inventory within a quarter. Programs that treat it as a standing function — network egress review, API-key audits, procurement flags, employee attestation — keep finding them before an incident does.

From Our Reference Shelf — The Samsung ChatGPT Leak (2023)

In March 2023, Samsung's semiconductor division, which makes the chips that power phones and data centers, gave its engineers permission to use ChatGPT, the AI chatbot made by OpenAI that generates text responses to prompts, for work. Within twenty days, three employees had independently sent confidential material to the service.

One pasted proprietary source code from an internal measurement database to troubleshoot an error. A second pasted code for identifying which production-line chips were functional and which were defective. A third recorded an internal meeting, transcribed it using a consumer transcription app, and fed the transcript to ChatGPT to generate meeting minutes.

By May, Samsung had restricted generative AI tools on company devices and capped what employees could paste to them.

None of it was an attack. There was no breach, no intruder. Three employees were trying to work faster with a capable tool that existed outside any inventory Samsung maintained. What they pasted went to a third-party service that retains everything you send it unless you explicitly opt out, on servers Samsung neither controlled nor could delete from. The harm was visible only after the data had left.

Samsung did not have a ChatGPT problem. Samsung had an inventory problem that ChatGPT made expensive. The tool was new. The failure was old: a capable system working on sensitive data with no registration, no owner, and no one who had ever been asked to approve it. Every organization reading this is one motivated employee and one consumer login away from the same week.

Incident Artifact — #ai-discovery, Day 14

#ai-discovery · 14:02

NO GOVERNANCE INSTRUMENTATION

d.okafor: found one. cs-team ticket summarizer. not in the registry.

d.okafor: it's calling an external LLM on a personal API key. someone's own account.

priya.n: whose?

d.okafor: no owner field. no registry entry. no tier. it's just... running.

d.okafor: and it's pasting ticket bodies into the prompt. ticket bodies have customer PII.

priya.n: how long has it been up

d.okafor: first call in the logs is 4 months ago. ~9k tickets.

gateway.egress.log (filtered: unsanctioned_endpoint=true)

```
ts=2025-... method=POST host=api.<external-llm>.com
auth=user_pat:cs-rep-0447 agent=cs-ticket-summary(unregistered)
payload.preview="summarize ticket #88231: cardholder [REDACTED-PAN],
DOB [REDACTED], dispute re: ..." owner=(none) risk_tier=(none)
```

#ai-discovery, Day 14 – one unregistered agent, 9,000 tickets, four months, customer PII in every prompt.

Nine thousand tickets across four months, one personal API key, no owner, and customer data in every prompt. Not one control in the building had a reason to look, because as far as every system that governs agents was concerned, this agent did not exist. Daniel does not have a word for the pattern yet. He just knows he is going to find more of them.

How to Run This at Your Company — The 30-Day AI Discovery Sprint

You cannot govern AI systems you do not know exist. The ones doing the most damage are the ones nobody thought to report.

Week 1. Look where AI actually lives. Three places hold the real inventory, and none of them is the official list.

Your system logs (1): they show what is actually calling an AI model, regardless of whether anyone approved it.

Your vendor contracts (2) : many software providers have added AI features to products you already own, and those systems count.

Your bills (3) : cloud and AI subscription spend, broken down by team, will surface tools nobody disclosed.

Start there. The official registry is where you end up, not where you begin, because it only contains what people remembered to register.

Week 2. Count what is running, not what was approved. For every AI system you find, record five things: its name, its owner, its purpose, its permissions, and its current stage (pilot, production, or retired). Do not spend a month designing the perfect tracking template. A short list with five honest columns is more useful than an elegant system with fifty empty ones.

Week 3. Find the AI your employees brought in themselves. This is the step most organizations skip, and it is the one that causes the real damage. Check whether employees are using personal AI accounts on company data. Look at outbound traffic for connections to AI services the company never approved. Ask managers directly what tools their teams are actually using to get work done. You are not looking to punish anyone. You are finding the tools that spread because the approved option required a formal request and the deadline was Tuesday. Expect this step to grow your count more than any other.

Week 4. Rank the risk and name an owner. Give every AI system a risk level and one named human accountable for it. People will push back on being named. That pushback is the most valuable information you will collect all month, because the systems nobody is willing to sign for are exactly the ones that either need a signer fast or need to be turned off.

The deliverable at day thirty is not a finished governance program. It is a number you trust and a list with owners. Everything that follows depends on both.

Maturity Placement

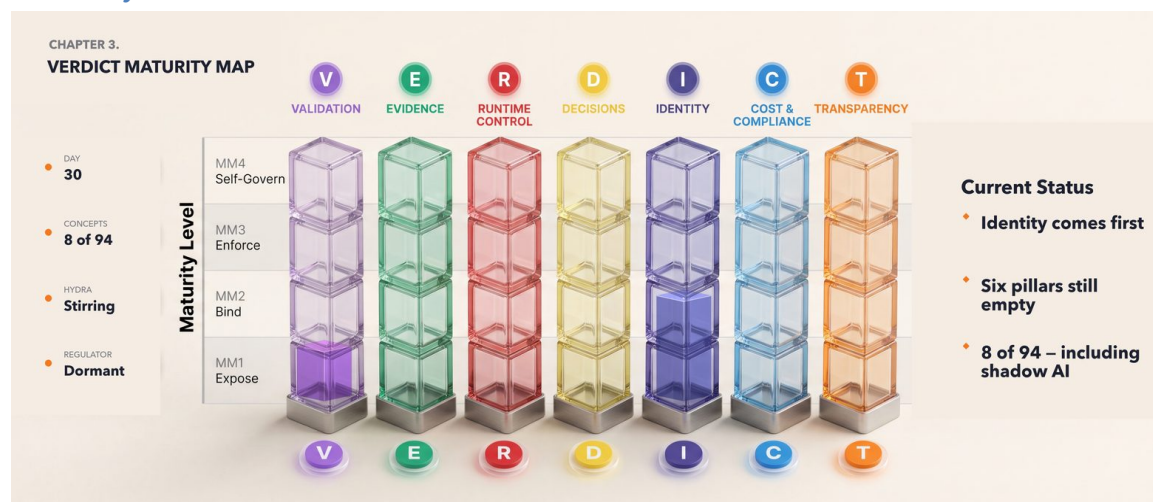


Figure 3.9 – Chapter 3 Maturity Map.

Velgrand, Day 30. Discovery in motion. The count is climbing, not settled. Identity not yet assigned – the agents have been found, not yet named. Level 1.

STORY SECTION B

Scene 4 – The count

Day twenty-one. The number on the whiteboard is two hundred and eleven. Three weeks ago, no one in the building believed the count would reach two hundred. It is still moving.

The forgotten agents no longer surprise Jane: the vendor feature nobody switched off, the integration built by someone who left in 2023. What catches her now is a different column Daniel keeps off to the side, in a marker that is running dry. The column is labeled Personal. It tracks agents tied to a single employee's own account, running on Velgrand's data because the company-approved tool was slower than the work.

Devika has been appearing in the conference room in the evenings, standing near the board, saying almost nothing. She has not offered support and Jane has not asked for any. But she is here, after hours, reading the columns.

Jane looks at the two lists, the forgotten and the personal. These are not shadow tools. A shadow spreadsheet stays where you put it. These agents read customer data and write to systems, and the count she has been building is not the achievement. Not one of these systems has a name, an owner, or any constraint on what it can do.

"The number isn't the problem," she says.

Devika answers from the wall. It is the first full sentence she has offered all week. "No. The number was never the problem." She caps the dry marker. "I've been running asset inventories for twelve years in security. You know what every single one has in common? Rows with no owner. The number isn't what gets you. It's the blank fields."

Scene 5 – The pushback

Day twenty-eight. Jane brings the count to Marcus before someone else does.

Three hundred and one. She puts the number in front of him, the forgotten and the personal both. Marcus reads it. He does not flinch, which she notes. Then he reframes it.

“This is shadow IT, the practice of employees using tools without IT’s knowledge,” he says. “It’s a real problem. We’ve had it for twenty years: people using tools we didn’t sanction. The answer is a capability question, not a tool question. We give people better sanctioned options and the shadow stuff goes away.” He slides the page back.

“You’ve done a good audit. But you’re describing a problem we’ve managed for twenty years, and I’m not going to reorganize my engineering org around an inventory.”

He stands. “Send me the list. I’ll have my team look at the company-approved-tooling gap.”

He walks out having conceded nothing and reframed everything. Shadow IT is a problem you manage. What Jane is looking at is a problem you govern. She has the count. He has the framing. The count was the easy part.

END-OF-CHAPTER BEAT

A push notification, late, on Jane’s phone.

AI-DISCOVERY: new instance found — count at 312.



THE VERDICT OPERATING MAP

V Validation E Evidence R Runtime Control D Decisions I Identity C Compliance T Transparency

