

KNOW YOUR AGENT

A Discipline for Financial Institutions in the Age of AI

Stephen Bishop

Tony del Fierro

with Atlas

PREVIEW EDITION

Introduction and Chapter 1

kyabook.com

Introduction

The 8-K I keep coming back to wasn't filed by a major. It was filed by a community bank, a small one whose name most readers wouldn't recognize.

I read it twice. I knew people who could have written it.

On May 12, 2026, the institution disclosed an AI-related incident. An employee had been using an unauthorized public AI tool to do parts of the daily work, and customer information had ended up inside the tool's interactions. The public reporting around the filing suggests this had been going on for some period of time before the institution caught it. Like most workarounds, this one likely had a believable origin story. The internal tools were probably slower than the public assistant. The policy presumably said don't do this. Whatever supervision was supposed to catch the workaround apparently didn't, until disclosure became unavoidable.

From everything I can read in the filing and the surrounding reporting, the institution wasn't hacked, and no vendor was breached. The institution had what it described as a written AI policy. Employees had been told about it. None of that mattered enough to prevent the disclosure.

The institution's exposure wasn't a system. It was a workflow.

It was also, as far as I can tell, the first U.S. financial institution to self-report an AI-related customer data exposure at that scale, and the word self-report is what keeps me coming back to the filing. How many other institutions had the same kind of incident and didn't disclose it? How many have it happening right now and don't know? How many will only find out when a customer notices, when an examiner asks, when a chat history surfaces somewhere it shouldn't have?

Those questions don't have a discipline behind them yet. Banking has built disciplines for everything else.

KYC tells a financial institution who its customers are. KYB tells it who the businesses on the other side of a relationship are. AML tells it how to recognize money trying to launder itself through the institution. BSA tells it how to write down what it saw. Each of those started as a problem nobody had named, became a regulatory expectation, and eventually became a discipline that shaped how banking is done. Underwriting, supervision, vendor management, model risk management. Each was a similar build, in a different shape.

What ties them together is that they are disciplines. They have their own vocabulary, their own artifacts, their own examination history, and their own way of failing and being corrected. Policies and tools sit inside each discipline, the way procedures and checklists sit inside KYC. An institution can pass a KYC audit or fail one, and the discipline is what gets tested, not the policy.

The book you're reading makes a simple argument: agents are the next discipline.

They are a new kind of party the institution interacts with, gets influenced by, and ends up trusting through repeated use. The existing disciplines were built for customers, businesses, transactions, vendors, and software. Agents borrow features from all of those and fit cleanly into none of them. Trust accumulates, and practical authority follows, so by the time anyone formalizes a question, the answer has already been getting answered through behavior.

That is why institutions need KYA. Know Your Agent. The discipline that does for agents what KYC does for customers, what KYB does for businesses, what AML does for money, and what BSA does for the record.

The examiners will ask you about your KYA. The prudent institutions will already have an answer. The rest will be the next 8-K.

* * *

That 8-K didn't come out of nowhere. In early 2023, the largest banks responded to ChatGPT the way institutions always respond to a new tool that scares them: they banned it. JPMorgan, Bank of America, Citigroup, Deutsche Bank, Wells Fargo, Goldman Sachs, all in the same week and with the same posture. Three years later, surveys show nearly half of workers admit they've used AI tools at work after the company said no, and a meaningful share have entered sensitive or proprietary data while doing it. The ban was the policy, but the behavior was something else.

The largest banks figured out earlier than most that the workaround was where the problem lived. Morgan Stanley built a walled-garden assistant inside its own environment, with prompts and responses confined to systems already governed by its compliance and data protection frameworks. Goldman moved from a blanket ban to internal AI deployment in its own environment by 2025. How effective that internal governance has been isn't fully visible from the outside. What was clear is that both institutions decided the sanctioned path had to be easier than the workaround.

Most institutions don't have the resources to build a walled-garden AI environment. Most institutions also don't get to decide whether AI enters their institution. The major core processors are bringing AI in through the infrastructure they operate. FIS has signed with Anthropic. Fiserv has partnered with OpenAI. The cores serve thousands of community banks and credit unions that don't have the leverage to demand specific governance terms from their core vendor. AI is arriving through the infrastructure the institution uses every day, whether or not the institution's AI committee has reached a decision on its own posture. Governance can't be opt-in when the AI is arriving through infrastructure choices made above the institution.

In April 2026, the OCC, the Federal Reserve, and the FDIC issued revised interagency model risk management guidance that explicitly extends the existing framework to generative and agentic AI, including chatbots and agents used for customer interactions and internal decisions. That guidance closes the "we didn't procure it" loophole, and it cuts both ways: any AI system influencing outcomes inside the institution is a model that needs governance, regardless of whether it was centrally procured or accessed through someone's personal browser tab.

The US isn't writing this on a blank page. The EU AI Act's general-purpose AI provisions took effect in August 2025, and the high-risk provisions land in August 2026. Revisions to PSD3 are converging on agent-mediated transactions. European institutions will be operating under agent-governance frameworks twelve to eighteen months before US regulators land their full version. Whichever European institution builds the discipline first becomes a reference case the US prudential regulators will read closely. The patterns set there will travel.

And the sanctioned AI deployments aren't safe either. A January 2026 security study tested twenty-four AI banking assistants using prompt injection and related red-team techniques across a range of deployment configurations. Every one of them was exploitable. The leak rates ran from one percent to sixty-four percent depending on configuration. One study is one study, and the methodology will be debated, but the directional finding, that polished vendor security claims don't survive adversarial testing under load, is the part the institution should not assume away. The vendor security report wasn't the test.

The illusion of safety is its own surface. Consumer AI products are marketed as safe for personal data. ChatGPT now encourages users to upload financial statements for personal financial management. The vendor language emphasizes privacy commitments and data handling promises. Consumers read those promises and act on them. So do employees, who are also consumers when they're not at work. When

the same person who uploads her tax returns to ChatGPT on Sunday opens her bank's complaint queue on Monday, she carries the same trust assumption into the workflow. Marketed safety isn't governable safety, and the institution that doesn't know the difference is going to find out the hard way.

* * *

The May incident is one shape of the problem. It's the employee surface, where the path of least resistance leads around the policy and the policy never catches up to the workflow. There are others.

Vendor-supplied AI is its own surface. An institution licenses a system, runs it inside its own environment, and treats it as governed because procurement signed a contract. The contract didn't test the model, the vendor's data flow runs through subprocessors the institution never reviewed, and the model was fine-tuned on data the institution can't audit. The institution thinks it has one vendor relationship. It actually has a chain.

The chain is wider than most vendor management programs are tracking. The core processor is one layer. The middleware vendors layered on top of the core, loan origination, member services platforms, fraud and AML tools, are another, each shipping AI modules in regular product releases without reopening the third-party review. The fintech partners absorbing AI features the EDD didn't reach are a third. The processors and payment rails introducing agent-initiated transaction support without changing the issuer agreement are a fourth. The customer is the fifth. Each layer carries an AI dimension that traditional diligence doesn't reach, and each layer is a chapter waiting to happen later in this book.

Customer-facing AI is another. Financial institutions are putting chatbots and agents into servicing, fraud, complaint handling, and account opening. Every one of those interactions is a surface where the institution is being trusted, the customer or member is being shaped, and the AI is operating in the middle of a regulated relationship.

And customers and members aren't starting to bring their own AI to the institution. They're already there.

By the end of 2025, nearly half of US adults had used a conversational AI assistant, and forty-six percent of Americans reported using AI for personal finances. An early 2026 industry tracker found that the mainstream-user share of AI for finance and banking doubled in a single month, from fourteen percent to twenty-eight percent.

Two hundred million people query ChatGPT on personal finance every month, and ChatGPT now connects to customer bank accounts via Plaid. Forty-two percent of small businesses and forty-five percent of medium-sized firms report watching AI agents make purchases on their behalf. McKinsey estimates three to five trillion dollars in global retail spend could flow through agentic channels by 2030, and Bain estimates fifteen to twenty-five percent of e-commerce in the same window. The institution's own transaction file already shows it, even if no one is reading the file as AI exposure data yet.

The agents arriving from the customer side are acting on behalf of a person to open an account, dispute a charge, move money, or file a complaint. Identity, authentication, and authorization frameworks were built for humans. A customer's agent isn't a human.

This isn't third-party risk management. It's all-party risk management: the institution, its employees, its vendors, its vendors' vendors, its customers, and its customers' agents. Every party that touches the institution's operating surface now has an AI dimension, and most of those dimensions aren't covered by the controls the institution already has.

There is a practical audit any institution can run today without buying anything new. ACH outbound to ChatGPT, OpenAI, Anthropic, Perplexity, or Google AI subscriptions, likely one to three percent of active members already. Bill pay and card-on-file activity to the same merchants, likely double the ACH figure. Card transaction volume routed through Instant Checkout, Buy It In ChatGPT, Comet, or Rufus, which means reading merchant-initiated transaction flags most cores already expose and most institutions never look at. Plaid pulls originated from OpenAI or equivalent agentic clients, which every institution can already see today, even though most have no policy on whether to be alerted, alarmed, or indifferent. Dispute and complaint volume that begins with the phrase "the AI told me," a category nobody tracks as a category yet. Each of those queries returns a number the institution already has. Nobody is reading them as AI exposure data because no one has framed them that way. The data exists. The reading discipline doesn't.

* * *

Underneath every one of those surfaces is a question the institution walked into and never answered.

What is this thing entering? Not the network or the contract. The institution. The actual operating environment with its policy gaps, its inconsistent procedures, its

queues already under pressure, its teams that document decisions one way and perform them another, and the workflows where speed already outruns documentation and everyone knows it.

That is the governance gap this book is about.

The gap isn't a tooling gap or a vendor gap or a policy gap. Tools improve, vendors mature, and policies get rewritten. The gap is the distance between what the institution thinks it has approved and what's actually happening inside its workflows, between the language in the policy and the behavior on the floor, between the version of the institution that's documented and the version that's operating.

* * *

Financial institutions already know how to govern trust between parties they can identify, classify, and supervise. What they don't yet have is a framework for governing the ones they can't.

The word matters. These are active participants in the workflow. They summarize, recommend, draft, route, prioritize, monitor, and in some settings act. They get trusted through repeated use, accumulating practical authority inside live workflows before anyone has said out loud what they're allowed to influence.

An institution may think an agent is only drafting language, while the front line starts treating that draft as the default answer. An institution may think an agent is only summarizing a file, while the summary becomes the frame through which the next reviewer sees the issue. An institution approves a pilot in one queue, then three months later the same pattern of use is spreading into compliance review, servicing, and escalation handling. Nobody approved that. It just happened because the thing was helpful.

Once that starts, the institution has moved past experimenting with a tool. It is teaching itself how much machine influence it can absorb before governance catches up.

* * *

This pressure shows up in exactly the places financial institutions can least afford to govern casually: complaint management, fraud case handling, servicing, operations exception review, policy interpretation, treasury onboarding, quality assurance, and internal research. These are the workflows where customer treatment, management accountability, and examiner scrutiny all meet.

Most institutions are approaching this the way they've always approached new software: choose a use case, pick a vendor, review the contract, limit access, launch carefully. That playbook worked when the product role was stable and the practical authority stayed easy to describe. It doesn't work for agents.

Agents absorb context and shape workflow. They reveal weaknesses that were already in the operating environment, then make those weaknesses move faster. If the institution's own ground is thin, the agent will expose it. Fixing the ground is the institution's work.

Before an institution can govern agents well, it has to understand the context those agents are entering: policies, procedures, audit findings, control gaps, committee habits, exception patterns, workflow artifacts, training quality, vendor assumptions, and the unspoken operating norms that hold the place together. An institution that hasn't surfaced its own operating logic will struggle to judge what good agent governance should look like in its own environment.

That is the case for Know Your Agent.

The central question of this book is simple: what exactly is this system, what role is it playing, what authority does it carry in practice, and what evidence supports trusting it in that role?

The argument is equally simple. AI governance in financial services is coming whether institutions build it deliberately or get dragged into it under pressure. The institutions that do this well will be the ones that can describe what the system is, what it's allowed to do, who owns it, and what would cause the institution to slow down before habit outruns judgment. The best slogan won't substitute for that description.

This book moves in three parts. Part I explains why the old governance language breaks down across all the surfaces AI is entering. Part II shows how KYA works in practice across internal, vendor, and customer-facing agents: classification, verification, operating ownership, implementation. Part III looks ahead to board oversight, supervision, credentialing, and the trust infrastructure that machine-mediated financial services will require.

The work starts in Chapter 1, and KYA starts with KYE.

Chapter 1. Know Your Enterprise

I launched a podcast with my friend Phil Goldfeder from the American Fintech Council. We named it Bourbon and BaaS (short for banking-as-a-service). The premise was simple: we drink bourbon and talk about banking, in particular fintechs and sponsor banking. The first episode went live on a Tuesday. By the next morning Phil was texting me. People were asking about the show, it was working, we needed a logo.

Easy. I'll ask AI.

I gave it the prompt. We're doing a podcast, we drink bourbon, we talk about banking-as-a-service, we call it Bourbon and BaaS, and I need a logo.

The image came back. It had sheep on it.

It took me a second to figure out what happened. The AI heard "Baaahhhs," not "BaaS." It didn't know I was a banker, it didn't know Phil was at AFC, it didn't know the show was about banking-as-a-service, and it didn't know what "BaaS" meant in context. So it grabbed the closest sound it had heard before, built an image around it, and handed me a logo full of sheep.

The tool worked. It produced exactly what I asked for, given what it knew, but I hadn't given it what it needed.

* * *

I tell that story because it's a small, funny version of a much bigger problem.

AI systems don't usually fail because of technical issues, they fail because of context issues. The model works, the vendor delivers, the pilot passes, and then the system lands inside an institution whose policies, procedures, exception patterns, and unwritten conventions weren't built for a participant that runs faster than the institution's own documentation of itself.

The model is the seed. The institution is the soil it lands in. The seed grows whatever the soil can grow.

The sheep on my logo were a contained mistake. I could see them, I could throw the image away, and I could go back and prompt better.

When the same kind of context mismatch happens inside a financial institution, nobody sees the sheep. The AI drafts a customer response, a reviewer skims it and clicks send, and the draft becomes the customer record. In account servicing, that draft is the response that goes out to the customer. In fraud investigations, it's the case summary that gets attached to the file. In complaint handling, it's the framing the issue gets investigated under. Three different departments, three different unwritten conventions about what counts as the record, and the AI didn't break any of them. It just made all three faster, more legible, and more invisible.

The institution is the variable, and the agent just makes the variable visible.

That is why the work starts earlier than most institutions think. Before an institution can govern an agent well, it has to understand itself well enough to know what governance would even mean.

This discipline has a name: Know Your Enterprise.

Know Your Enterprise

You have to know your enterprise before you can understand how it operates, how AI is affecting it, or how best to govern that AI. Until you know what the institution is, the other questions don't have honest answers.

That is what Know Your Enterprise means.

KYE is the institution reading its own policies, procedures, board minutes, audit findings, exception logs, training records, and committee reports the way it reads a new customer's KYC file. The assumption running through it is that anything important that isn't written down still matters, and anything written down still has to be tested against behavior.

KYE isn't a data inventory or a control catalog, it's enterprise self-awareness. It's the institution's own answer to what it is, how it actually works, where its documented practices and its lived practices come apart, and where it carries authority that no one wrote down.

An institution that hasn't done KYE will struggle to do KYA. The agent's authority will be measured against an environment the institution doesn't fully understand, and the trust question will be answered against a worldview that hasn't been surfaced.

That is the deeper claim of this book: KYA starts before the tool arrives, and it starts with KYE.

Preparing the soil

If KYE is the discipline, soil is the metaphor for what the discipline reads, and the gap between the documented institution and the operating one is what the discipline reveals. The agent isn't software in the legacy sense. It's a seed being planted into soil the institution has cultivated, ignored, drifted from, and worked around for years. The soil decides what grows.

Strong soil grows strong outcomes. Weak soil still grows something, just not what management thought it was growing.

That's what the sheep were. The AI grew exactly what it could grow from the context I gave it. I'd handed it weak soil, and the output reflected the soil, not the seed.

Financial institutions are doing the same thing every day, at a much larger scale, with much higher consequences. The agent gets deployed into a workflow whose policies are inconsistent, whose procedures vary team to team, whose exception handling depends on experienced people who can't always explain their logic, whose committee reporting smooths over recurring friction because everyone has learned to live around the weak spots.

That isn't unusual, it's normal institutional life. The problem is that agents don't enter institutions neutrally, and they amplify what's already there. In a well-run workflow, they may increase speed without breaking discipline. In a weakly understood workflow, they turn partial clarity into faster partial clarity, which isn't the same thing as control.

That's why some deployments feel fine in the demo and unstable in production. The institution thinks it's evaluating the system, but the deployment is also stress-testing the institution.

Surfacing the institution

KYE starts with artifacts the institution already has: policies, procedures, training materials, audit findings, issue logs, exception patterns, quality reviews, committee minutes, escalation paths, workflow maps, and vendor assumptions. Most institutions never read these together. Compliance owns one stack, operations owns another, audit owns a third, and the artifacts only get pulled into one room when

something has already gone wrong. KYE is the discipline of reading them as a single document about the institution's own behavior, before something goes wrong.

Taken separately, those artifacts feel administrative. Taken together, they form a worldview: how the institution thinks work should happen, where that view is strong, and where it's already compromised by drift, inconsistency, or informal habit.

Underneath the artifact list are four things KYE has to surface honestly. Policies are what the institution has committed to. They matter because the agent will operate inside the policy environment regardless of whether that environment matches the institution's actual practice. Procedures are how the work is supposed to happen step by step. They matter because the agent will absorb the procedure as a template, which means a procedure that is half-followed becomes a template that gets half-followed faster. Operating reality is the version of the institution that shows up in queues, exceptions, and informal habits. It matters because the agent will work inside this version, not the policy version. Decision authority is the question of who is allowed to decide what, and at what level. It matters because the agent will quietly accumulate decision authority through use, and the institution that hasn't surfaced its own authority map will not notice the accumulation until something forces the question.

That compromise is where most institutions actually live. Procedure says one thing, and the exception log records the other thing happening in most of the cases. A committee minute records a decision to require dual signoff, and three months later, two teams have quietly stopped doing it and nobody escalated. Training materials still describe last year's screens, and a manager tells the new hire on day three to ignore most of it. Vendor packets promise that customer data won't be used for training, then bury an exception two paragraphs later that nobody read. None of this is fraud or neglect. It's the normal way institutions drift between the version of themselves they document and the version of themselves they operate.

The worldview matters because KYA isn't just a method for inspecting the agent. It's a method for testing whether the institution understands its own operating logic well enough to place the agent responsibly. The agent will land inside the drifted version, not the documented one. It will get trusted, modified, expanded, and absorbed into the lived workflow regardless of what the policy says. Knowing the drift before the agent arrives is the difference between governing the system and being governed by it.

KYE is the test the institution runs on itself before asking the agent to pass any test of its own.

What a KYE pass looks like in practice is less mysterious than it sounds. A working group with representation from compliance, operations, audit, the business line, and a leader who can convene them pulls the artifacts into one room, reads them together against three workflows that matter most to the institution, and writes down what it finds. The deliverable is a short operating-reality memo: where the documented institution matches the operating one, where it doesn't, and what governance the institution would need to keep both versions honest as agents get introduced. That memo isn't the answer to everything. It is the first piece of evidence that the institution has seen itself. The full mechanism of the KYE pass, the working session structure, the artifact list, the operating-reality memo template, lives in Chapter 14's starting plan.

Old software instincts don't carry

Earlier software cycles trained institutions to look for seriousness in recognizable places. The company looks mature, the packet is clean, the controls are documented, the product category is familiar, and the workflow impact seems bounded. Those instincts were never perfect, but they often got the institution far enough.

Agents weaken that comfort.

The useful system may arrive with a thinner package, or from a provider whose polish lags its functionality, or configured internally in ways that change its operating role more than the vendor packet suggests, or it may become influential because employees find it helpful rather than because anyone intentionally granted it more authority. The institution can no longer assume that a familiar procurement rhythm will produce a truthful governance picture.

That procurement rhythm has a name: third-party risk management (TPRM). TPRM was designed for relationships the institution chose, contracted with, and reviewed annually. Agents extend past that boundary, which is why this book argues for all-party risk management, or APRM, as the broader frame. The full case for that argument comes later. What matters here is that TPRM doesn't reach the parts of the institution where agents actually take hold. This chapter is about the part TPRM never covered: the institution itself.

That's why KYE matters. It forces the institution to do the harder work of understanding its own environment before the tool becomes useful enough to outrun the original approval story.

Readiness isn't interest

When institutions say they want to be ready for AI, they often mean they want a policy, a vendor process, and maybe a steering committee. Those things help, but they aren't the full answer.

Readiness means the institution can explain the conditions into which the agent is being introduced. It means management understands the workflow well enough to know where machine influence might accumulate quietly. It means the institution can identify where existing controls are honest, where they're ceremonial, and where the process already relies on judgment that has never been made explicit. It means the institution is honest about the distance between written procedure and lived practice.

That kind of honesty isn't glamorous, but it is where better governance begins.

Most institutions are still mistaking interest for readiness. They have policies that say the right things, they have steering groups, they have vendor packets, and they have enthusiasm from the business balanced by caution from the control functions. What they don't yet have is an operating discipline strong enough to keep pace once the system starts becoming useful.

A more sophisticated version of the same mistake looks like prudence. The institution picks a single workflow and deploys a chatbot inside it. The pilot has limited scope, contained risk, clear ownership, and an easy rollback if things go wrong. The framing sounds responsible: the institution isn't betting itself on AI, it's starting small.

That approach is wrong: being ready for one workflow isn't being ready for AI. The chatbot doesn't stay in the workflow it was deployed for. The front line finds adjacent uses, demand pulls it into queues nobody approved it for, and three months later the institution is operating an agent it never officially expanded. The institution that hadn't done KYE before deployment is no closer to having done it afterward. It has just compressed the readiness question into one queue and called the absence of an answer "careful adoption."

Usefulness creates drift. The more the tool helps, the more pressure the institution feels to widen the role. The more pressure it feels to widen the role, the thinner the original approval language starts to sound. By the time management realizes the institution is relying on the system more heavily than planned, the workflow may already have adapted around it.

That isn't readiness, it's governance lag.

* * *

The rest of this book is built on a premise. Financial institutions don't need more abstract AI enthusiasm. They need a clearer way to understand when software starts behaving like an operating participant, what kinds of trust that requires, and how to keep governance aligned before speed turns into dependence.

Back to the logo. The tool worked for me. It produced exactly what I asked for, given what it knew. I hadn't given it what it needed. That sentence is the whole problem, scaled. The institution that hasn't done KYE is the institution that hasn't given the tool what it needs to work for the institution rather than against it. The sheep show up in different forms inside every workflow the institution hasn't surfaced honestly.

KYA is the discipline that does that, and KYE is where it starts.