SYMPHONY

Securing the Next Frontier: Agent-to-Agent Communication in Financial Services

The case for agentic AI in financial services is well made: faster synthesis, lower operational cost, scalable analytical capacity. Despite such breakthrough progress, the case for good governance has considerably lagged behind.

As autonomous AI workflows move from internal experimentation into production environments, they create a new class of compliance and security challenges that existing frameworks were not designed to address. This paper sets out what those challenges look like in practice and why early movers on governance infrastructure will hold a structural advantage as regulatory expectations evolve.

The history of technology adoption in institutional finance follows a familiar pattern. *A new capability typically emerges in the consumer or enterprise software sectors. Early adopters conduct controlled pilot programs to test their potential. Regulators then start to raise questions about its implementation. A significant incident occurs. Suddenly, what was once seen as a discretionary investment becomes an urgent necessity.*

Chief technology officers, compliance and risk leaders at banks, asset and wealth management must now confront a critical question: Will agent-to-agent AI workflows follow a similar trajectory to past technological shifts?

The stakes are higher than they have been for any previous technology cycle. The convergence of Large Language Models (LLMs), the Model Context Protocol (MCP) and multi-agent orchestration frameworks is driving a structural shift in how decisions are made, how data moves and ultimately who, or what, is accountable when something goes wrong.

From internal tooling to autonomous workflows

Over the past three years, the deployment of AI in institutional finance has been characterized by careful and limited application. This has been almost exclusively internal: algorithmic trading strategies informed by machine learning, LLM-assisted analysis of earnings transcripts and regulatory filings, automated extraction of structured data from complex legal documents and generative tools to accelerate the production of pitch materials and research summaries.

These deployments share a common characteristic. A human professional remains in the loop at every consequential decision point. The AI augments, but it does not act independently.

The boundary between human decision-maker and automated execution is, however being systematically dissolved.

The emergence of MCP as an open standard for how AI models interact with external tools, databases and services has made it materially easier to chain together specialized AI agents into coherent workflows. Where a trader or sales assistant once spent hours coordinating trade details across chat, order management systems, risk checks and booking platforms, an orchestrated set of agents can now manage that workflow end-to-end in minutes, without manual handoffs between each step.

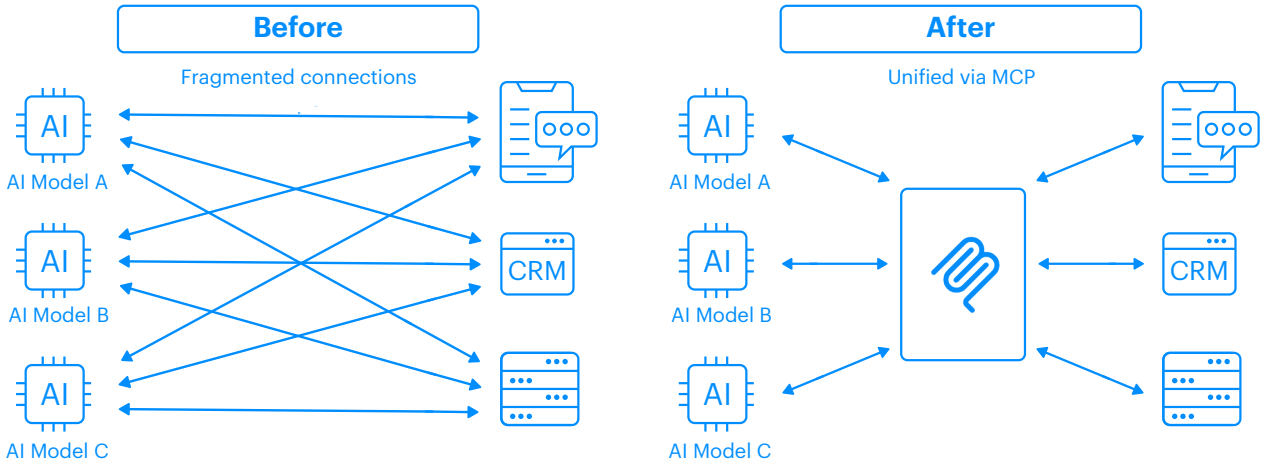
Spotlight

JP Morgan has begun deploying agentic AI for complex multi-step employee tasks. Its chief analytics officer, Derek Waldron, gave CNBC the first outside demonstration of the platform, showing it produce an investment banking deck in approximately 30 seconds, work that would previously have taken a team of junior bankers hours. The bank's LLM Suite is now accessible to roughly 250,000 employees, with around half using it daily. The stated end-state is striking in its ambition: every employee with a personalized AI assistant, every back-office process run by agents, every client interaction curated by an AI concierge.

A quick guide

What is Model Context Protocol (MCP)?

MCP provides a standardized interface for AI models to access external data and tools, reducing the need for custom point-to-point integrations. Instead, teams can build reusable connectors that link AI to core systems like local files, CRM platforms and messaging tools. This enables agents to operate with richer context and controlled access while governance, memory and permissions are managed by the surrounding infrastructure, making it suitable for regulated environments.



Before MCP, each AI model required its own custom integration. A single interface now connects AI to files, CRMs, databases and messaging tools with much less complexity.

What MCP changes for regulated firms:

- **Deployment speed creates governance pressure.** MCP reduces the time required to set up and develop AI solutions. For regulated firms, that speed advantage only holds if governance frameworks keep pace; otherwise, faster deployment means faster accumulation of unaudited risk.

- **Dynamic tool access demands dynamic oversight.** Agents that can discover and invoke tools at runtime are more capable and harder to supervise. Firms need surveillance controls that are as dynamic as the agents they are governing.

- **Modular architectures require modular accountability.** Composing AI workflows from specialized components is operationally powerful. It also distributes accountability across multiple systems, multiple groups and vendors, making clear ownership of each component a compliance requirement, not just good practice.

The rise of MCP has also accelerated the development of Agent-to-Agent (A2A) workflows.

Unlike single-agent or rule-based systems, multi-agent setups rely on specialized agents, each focused on a specific task or domain. These agents collaborate, sharing context and handing off work, much like a team of experts, enabling better problem decomposition, adaptability and more robust outcomes.

MCP and A2A are complementary: MCP standardizes how agents connect to tools and data (e.g. APIs, files, systems), while A2A enables agents to discover, communicate and coordinate with each other. Together, they form an orchestration layer where MCP handles execution and A2A drives collaboration and delegation.

This model is gaining strong interest from enterprise buyers, particularly in financial services. With high demands for data synthesis and limited analytical capacity, the sector is well-positioned to adopt these production-ready multi-agent patterns quickly.

A quick guide

What is an Agent-to-Agent Protocol (A2A)?

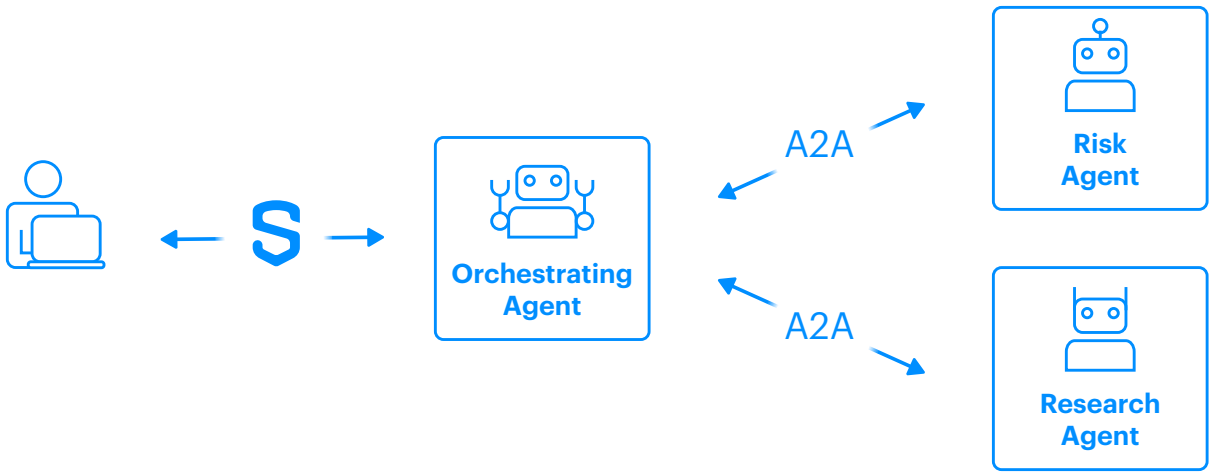
A2A is a standardized protocol enabling multiple independent AI agents to effectively communicate, collaborate and validate outputs. Instead of one large model, A2A distributes responsibilities across specialized agents, forming a cohesive, distributed and self-managing network. This approach boosts efficiency, flexibility and adaptability for complex workflows.

Spotlight

Morgan Stanley's Debrief tool rolled out to wealth management advisors uses OpenAI to generate meeting notes, draft follow-up emails and log call information directly into Salesforce with client consent. Nearly all of the firm's financial advisor teams have adopted some form of AI assistant. The firm's head of AI, Jeff McMillan, has described the next phase explicitly: AI serving as an "efficiency-enhancing interaction layer" sitting between colleagues and execution systems, CRMs, reporting tools and risk analysis platforms. That is, by his own description, a multi-agent architecture in the making.

Agent-to-Agent Protocol

An advisor invokes an agent to re-evaluate a portfolio allocation. The orchestrating agent routes the request to a research agent, which recommends increasing exposure to gold, while simultaneously engaging a risk agent to check exposure limits. Each agent accesses only the data it is permitted to see and all interactions are logged for audit. An A2A protocol enables these agents to coordinate seamlessly, operating as a chain of specialized assistants.



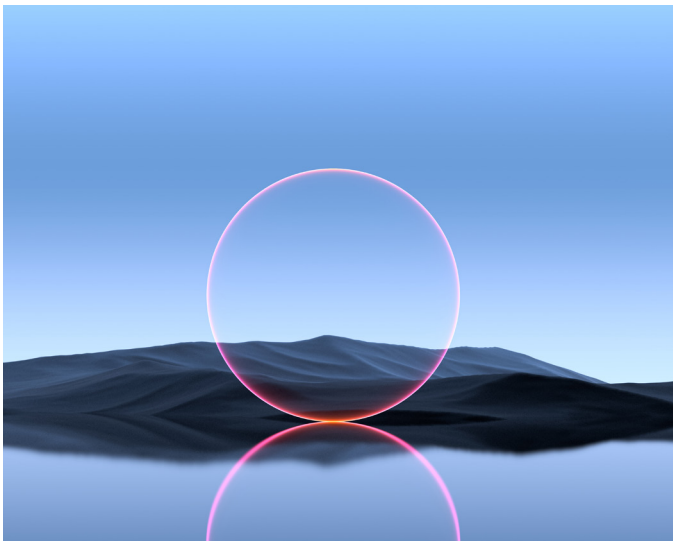
The new AI risk surface

Executives who have navigated the security and compliance dimensions of earlier technology transitions, from cloud migration to mobile device management and open banking API frameworks will recognize a familiar dynamic. Each time a new integration layer is introduced, the attack surface expands and the points of potential failure multiply. Agent-to-agent workflows change the conventional cybersecurity frameworks and introduce several categories of risk that deserve serious boardroom attention.

Trust and identity In multi-agent environments, reliability is impossible without verification for both human and AI participants. Establishing a verified trust network is essential before any task begins; agents must discover each other, authenticate identities and define clear operational boundaries. In the absence of such a foundation, the entire workflow is compromised. Vulnerabilities include agent spoofing, unauthorized privilege escalation by compromised participants and the structural risk of lateral movement across interconnected systems.

Prompt injection and manipulation The inputs to AI agents are not always controlled. In a workflow where one agent's output becomes another agent's input, an attacker who can influence the data that the first agent processes, through a poisoned document, a manipulated market data feed, or a compromised external API, can potentially direct downstream agents to take actions that were never intended. Security researchers have demonstrated proof-of-concept attacks of this kind against publicly available agent frameworks. Financial institutions, whose workflows regularly consume external data, face an elevated version of this risk.

Excessive agency The governance challenge that regulators are least equipped to address, and that financial institutions are least prepared to manage, is the question of what happens when an AI agent has been granted more access than it needs. If an agent responsible for summarizing client communications also has write access to a trading system, because that was the path of least resistance during a rushed integration, the consequences of a misconfiguration, a manipulation, or an unexpected emergent behaviour could be severe. The principle of least privilege, well established in cybersecurity, has not yet been systematically applied to AI agent design.



Spotlight

Anthropic documented in late 2025 what it assessed with high confidence to be a state-sponsored group that manipulated its Claude Code tool to attempt infiltration of roughly thirty global targets, including financial institutions. The attackers used AI's agentic capabilities to execute the cyberattack itself, with AI performing an estimated 80–90% of the campaign, requiring human intervention at only four to six decision points per operation. The attack used the MCP as part of its execution infrastructure. The attack demonstrated that the same autonomous, low-human-touch properties that make agentic AI efficient also make it an attractive execution vehicle for adversaries.

Transparency and auditability Financial regulators on both sides of the Atlantic have been explicit: they expect firms to be able to explain how decisions affecting clients or markets were reached. The FCA's Consumer Duty requirements and the SEC's scrutiny of algorithmic trading all presuppose a level of traceability that many current agent architectures do not provide by default. When a multi-agent workflow makes a recommendation or executes an action, the audit trail needs to be as legible as the decision itself. This is uniquely difficult in multi-agent systems because no single agent "made" the decision as the decision was distributed.

Cascading failure Perhaps the most underappreciated operational risk is the systemic nature of multi-agent workflows. A single malfunctioning agent; one that begins producing degraded outputs because its underlying model has drifted, or because it has been fed corrupted data does not fail gracefully. It propagates its errors to every downstream agent that depends on it. In a high-frequency or latency-sensitive environment, the window between a failure and its detection can be extraordinarily narrow. The failure mode is less like a server going down which can be visible, detectable, recoverable. It's more like corrupted data propagating silently through a supply chain before anyone realizes the finished product is defective.

The regulatory horizon

The regulatory posture across the US, EU and UK is converging on a single expectation: that firms can demonstrate meaningful human accountability for AI-driven decisions, even when no human was directly involved in making them. That is a technically demanding requirement that most current agent architectures cannot meet.

The regulatory response to agentic AI is still forming:

The EU AI Act, which entered into force in August 2024 and will apply to high-risk AI systems from August 2026, explicitly addresses automated decision-making in financial services. Systems that influence credit decisions, insurance pricing, or investment advice are classified as high-risk. These capture obligations around transparency, human oversight and technical documentation that many current agent deployments would struggle to satisfy.

In the United Kingdom, The Bank of England's review of AI in financial services, published in 2024, noted that multi-agent systems present novel challenges for model validation and governance — challenges that existing model risk management frameworks were not designed to address.

In the Asia-Pacific region, Singapore has moved furthest and fastest. In November 2025, the Monetary Authority of Singapore (MAS) issued Guidelines on AI Risk Management covering all MAS-regulated institutions across the full AI lifecycle. The guidelines explicitly address AI agents, requiring board-level oversight and, where AI risk exposure is deemed material, a dedicated cross-functional governance committee. MAS is unambiguous that institutions cannot delegate governance obligations to third-party vendors: if an external model is embedded in a workflow, the regulated firm remains accountable for what it does. That principle directly contradicts how many institutions are currently approaching AI procurement. MAS's guidelines are likely to set the regional standard for multinationals operating across APAC.

In the United States, the SEC has pulled back from AI-specific rulemaking, but its 2025 examination priorities signal that AI governance will be assessed under existing frameworks, making internal discipline more, not less, important in the absence of prescriptive rules.

Regulators understand the efficiency arguments for AI adoption and none of the major supervisory bodies has moved to restrict it categorically. The expectation, rather, is that firms will deploy these capabilities within a governance architecture that provides the oversight, traceability and control that financial markets require. That governance architecture does not yet exist off the shelf. Building this architecture is the challenge facing the industry now.

What a secure agent infrastructure looks like

The good news for financial institutions navigating this landscape is that the security and compliance requirements for agent-to-agent workflows are not entirely novel. They build on principles that the industry has applied, in other contexts, for decades: end-to-end encryption of communications, immutable audit trails, granular access controls and cryptographically verifiable identities.

The challenge is that most enterprise AI infrastructure was not built with these requirements in mind. The major AI frontier platforms offer powerful capabilities, but their default configurations prioritize ease of deployment over the control and auditability that regulated institutions need. The result is a significant integration burden for the compliance, technology and risk teams who are expected to sign off on production deployments.

The more coherent approach and the one that a small number of financial institutions are beginning to adopt is to treat the secure communications and compliance layer as foundational infrastructure, rather than as a bolt-on to an AI deployment.

1. AI agents should be deployed as credentialed participants within a controlled communications environment, not as standalone services with direct access to enterprise systems.

2. End-to-end encryption must extend beyond data at rest and in transit to include data in use, supported by;

3. Trusted Execution Environments (TEEs) that provide cryptographic assurance workloads run on untampered infrastructure. Existing audit trail obligations long applied to human communications in regulated firms, must also be extended to cover every interaction within AI-mediated workflows.

The principle is straightforward, even if the implementation requires careful engineering. AI agents, like human professionals, should operate within an identity and access framework that knows who they are, what they are permitted to do and creates an immutable record of what they have actually done.

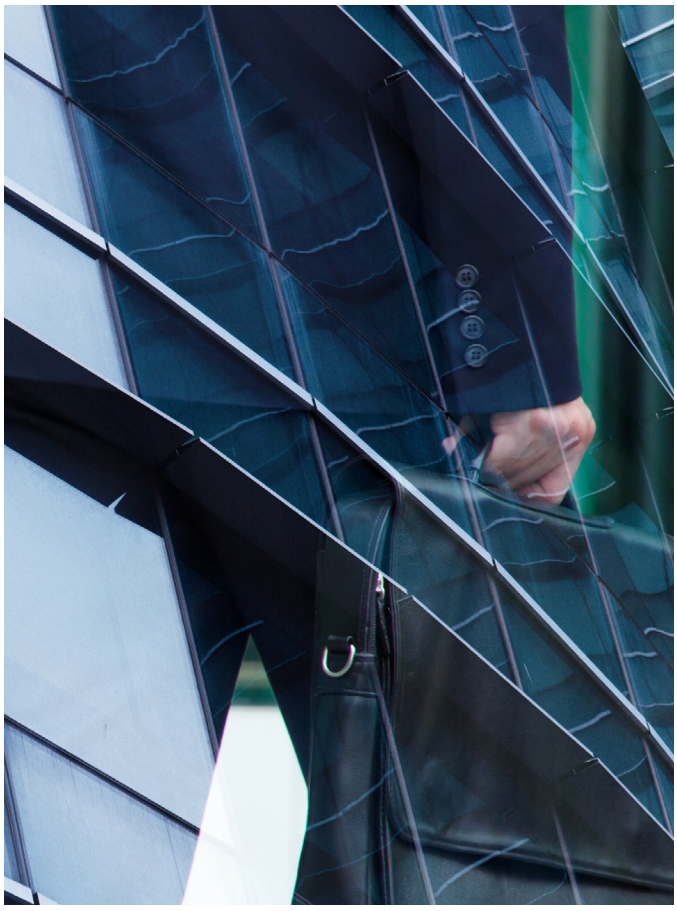
The compliance opportunity

For CxOs assessing the investment case, one point is critical: the governance architecture needed for safe agent deployment is more than cost. It is a competitive advantage.

Financial services firms that can credibly demonstrate to regulators, to institutional clients and to their own boards that their AI-driven workflows are auditable, controllable and consistent with their compliance obligations will be in a structurally stronger position as the regulatory environment tightens. Firms that have deployed AI capabilities in ways that cannot be explained or audited will face a costly retrofit or, worse, a supervisory intervention at a moment of their regulator's choosing rather than their own.

The early movers in cloud adoption in financial services understood this dynamic. Firms that embedded compliance into their architecture from the outset, rather than treating it as an afterthought, found that governance became a source of strength. It not only eased regulatory scrutiny but also enhanced client confidence and operational resilience.

The same logic applies here. The question is not whether to build governance infrastructure for agentic AI; firms need to do so now at a manageable cost, or risk much higher expenses later.



Looking ahead

The pace of agent-to-agent adoption in financial services is unlikely to be smooth and the timing of specific regulatory requirements remains uncertain. The direction however is resolute. Agentic AI capabilities will become more powerful, more widely deployed and more deeply integrated into core financial processes over the next three to five years. The firms that are building the governance infrastructure now are the ones that will be positioned to capture those capabilities and to do so in a way that their clients, their regulators and their own risk functions can stand behind.

The next challenge is equally significant: as AI agents begin to interact not just internally but with agents operated by counterparties, exchanges and infrastructure providers, the question of Federation - how trust and compliance standards are maintained across organizational boundaries will move to the top of the agenda. That is the subject we will examine in the next paper in this series.

About this series This piece is the second in a series of technical and strategic briefings examining the governance, security and regulatory dimensions of AI adoption in institutional finance. The series draws on contributions from Symphony's engineering, product, legal, and compliance teams.

Symphony's approach Symphony's AI offering is designed around the principle that secure, compliant, and verifiable AI infrastructure is not a constraint on capability. It is its foundation. By integrating AI agents as credentialed participants within Symphony's communications and compliance ecosystem, firms can deploy agent-to-agent workflows with the end-to-end encryption, audit trail integrity and access control architecture that regulated financial services demand. Symphony's Confidential Cloud, built on a TEE, extends these guarantees to AI workloads in active processing, providing cryptographically verifiable assurance that AI operations are authentic, isolated and tamper-evident.

To learn more about Symphony's AI offering and how it addresses the governance requirements explored in this paper:

Download the Solution Brief