

Agentic Artificial Intelligence for Autonomous Decision-Making in Finance

A decision-rights framework for trading, fraud detection, cash-flow forecasting,
risk early-warning, and customer interaction

Samuel Nyarangi

Independent researcher

Working paper · August 2026

Working paper. Conceptual synthesis. Not a trading, credit, or compliance manual.

Abstract

Financial institutions are moving from predictive models that score events to *agentic* systems that perceive a state, form a plan, call tools, and take actions that change books, cases, forecasts, and client relationships. Survey evidence shows executives intend to spend more on agents while remaining unwilling to trust them with high-stakes financial transactions. The research literature has expanded architectural experimentation faster than it has standardized evaluation, execution semantics, or reproducible artifacts.

This paper argues that the binding design problem is not which model is smartest, but who may decide what, under which mandate, with which evidence, and with which stop condition. It offers a decision-rights framework—perceive (R0), recommend (R1), execute inside limits (R2), update policy (R3), and coordinate other agents (R4)—embedded in a closed sense–plan–act–learn loop with a mandate gate. The framework is mapped onto trading, fraud and anti-money-laundering investigation, cash-flow forecasting, risk early-warning, and customer interaction. Perception and recommendation can usually be automated. Bounded execution is appropriate only when reversibility, loss caps, and replay are real. Policy updating and multi-agent coordination should remain rare, dual-controlled, and inventoried as models.

The contribution is conceptual rather than a new backtest. The practical implication is conservative: institutions should scale agentic *workflow* before they scale agentic *authority*.

Keywords: agentic AI; autonomous decision-making; algorithmic trading; fraud detection; cash-flow forecasting; early-warning systems; conversational banking; model risk; EU AI Act

JEL: G21, G32, C63, D81, K22

Scope

This paper does not estimate an average treatment effect, does not report a live-book backtest, and does not propose a structural forecast. Figures 1–4 and Tables 1–2 are **conceptual design maps**, not estimated indices. Author-reported returns and survey percentages are attributed to the cited sources and are not new results of this paper.

Contents

1	Introduction	3
2	Background and related work	4
2.1	From software agents to language agents	4
2.2	Agentic systems in trading and investment	4
2.3	Fraud, financial crime, and investigative work	5
2.4	Cash-flow forecasting and treasury	5

2.5	Risk early-warning	6
2.6	Customer interaction	6
2.7	The gap this paper fills	6
3	A decision-rights framework	7
3.1	What counts as an agent here	7
3.2	The agency spectrum	7
3.3	Decision rights R0–R4	7
3.4	The control loop	8
3.5	An institutional loss function	8
3.6	Human–agent residual responsibility	9
4	Mapping the framework to five financial functions	9
4.1	Trading	9
4.2	Fraud detection and financial-crime compliance	10
4.3	Cash-flow forecasting	11
4.4	Risk early-warning	11
4.5	Customer interaction	11
5	Architecture, evaluation, and failure modes	12
5.1	A common stack	12
5.2	What to measure	12
5.3	Characteristic failures	12
6	Governance and regulation	13
6.1	Agents are models, and more than models	13
6.2	European high-risk AI, operational resilience, and privacy	13
6.3	Conduct, markets, and crime	14
6.4	Organization	14
7	Research agenda	14
8	Conclusion	15

1. Introduction

For two decades, applied machine learning in finance has been a scoring technology. Models rank borrowers, flag transactions, estimate cash, classify news, and price residual risk. A human, a rules engine, or a separately coded execution stack then converts the score into an action. That division of labor is now under pressure. Large language models (LLMs) can generate intermediate reasoning, call tools, write and revise plans, and keep a working memory across steps. Surveys of this literature describe LLM-based autonomous agents as systems that combine a profile, memory, planning, and action interface rather than a single forward pass from features to a label¹. The ReAct pattern made the interleaving of thought and tool use explicit², and subsequent memory- and role-based designs imported that pattern into financial settings^{7–9}.

The economic lure is obvious. Trading, investigations, treasury, risk monitoring, and service are all sequential decision problems under incomplete information. Each already has humans who gather evidence, debate, act, and write a file. If an agent can do some of that loop at machine latency, the institution can compress cycle time, widen coverage, and leave specialists on the residual cases. Industry surveys record that intent. A May 2025 PwC survey of 300 senior executives reported that 88 percent of respondents said their team or function planned to raise AI-related budgets over the next twelve months because of agentic AI. The same survey found that stated trust collapsed for financial transactions, to 20 percent¹³. Capgemini’s research institute, drawing on 1,500 senior executives in fourteen countries, framed agents as a large potential source of revenue and cost savings *conditional on trust*¹⁴. S&P Global has likewise argued that agentic systems could help capital-market firms operate across fragmented workflows, not merely score isolated tasks¹⁵.

Those statements describe appetite and anxiety, not a settled operating model. In research, specialized trading agents now mimic desks: FinMem emphasizes layered memory and a trader “character”⁷; FinAgent adds multimodal inputs and tools⁸; TradingAgents assigns analyst, trader, and risk roles and lets them debate⁹. Adjacent work shows that even a general-purpose model can extract return-relevant information from news headlines⁶. In financial crime, vendors and research prototypes describe agents that assemble entity narratives, work low-risk alerts, and orchestrate case steps. Treasury commentary treats machine-learning cash forecasting as a core use case, while academic work on *generative* cash-flow agents remains thin. Early-warning research has a longer machine-learning pedigree in bank distress and borrower monitoring, and customer-service research has moved from scripted chatbots to LLM agents that must follow standard operating procedures.

What the field still lacks is a single account of *decision rights*. A trading agent that can send an order, a fraud agent that can file or close a case, a treasury agent that can move cash, a risk agent that can freeze a limit, and a service agent that can promise a refund are not the same object as a classifier with a high out-of-sample score. They change the institution’s position, its regulatory file, or its promise to a client. Once action is in the loop, model risk, operational resilience, market-conduct, privacy, and consumer-protection regimes all attach. The European Union’s Artificial Intelligence Act, in force since August 2024, already treats systems used to evaluate the creditworthiness of natural persons as high-risk, with an explicit exception for systems used to detect financial fraud^{26,27}. U.S. banking supervisors have, for fifteen years, treated quantitative methods that produce estimates used in decisions as models under SR 11-7²⁹; in 2026 the agencies issued updated interagency model-risk guidance that recasts rather than discards those principles³⁰.

The design problem is not one-sided. On one view, more autonomy raises coverage and cuts cycle time, which is why executives report planned budget increases^{13,14}. On the other view, the same autonomy raises false-action, herding, and conduct costs, which is why stated trust collapses for financial transactions¹³ and why supervisors treat decisioning models as inventory objects^{26,29}. Which force should dominate is not a software question. It is a mandate question: how far an agent may go before a human, a limit, or a kill

switch is required.

This paper’s claim is therefore organizational as well as technical. **Agentic AI is useful in finance to the extent that authority is decomposed, limited, logged, and challengeable.** The contribution is a portable rights ladder (R0–R4), a closed control loop with an explicit mandate gate (2), an institutional loss (3), and a domain map that keeps execution rare where actions are hard to reverse. Section 2 situates the argument. Section 3 states the formalization. Section 4 maps five functions. Section 5 treats architecture and failure. Section 6 treats governance. Section 7 lists open questions. Section 8 concludes.

The intended readers are researchers in financial machine learning and practitioners who must decide how far an agent may go before a human, a limit, or a kill switch is required.

2. Background and related work

2.1 From software agents to language agents

The idea that an agent is a system that perceives an environment and acts to achieve goals is older than the present wave of language models. Textbook treatments in artificial intelligence and multi-agent systems already distinguished reactive policies from deliberative planners and stressed that multiple agents create coordination, incentive, and equilibrium problems^{3,4}. What changed after 2022 is the interface. A sufficiently capable language model can write the intermediate reasoning that used to live only in a human’s head or in a separately engineered planner, and it can emit tool calls that look like actions. ReAct showed that interleaving verbal traces with actions improved both interactive decision tasks and the inspectability of the trajectory². Chain-of-thought prompting had already shown that forcing intermediate steps can raise reasoning accuracy⁵; tool-using variants then connected those steps to search, code, and application programming interfaces.

Wang and coauthors organized the subsequent explosion of LLM-agent papers around construction (profile, memory, planning, action) and evaluation¹. That taxonomy is a useful import into finance, because financial work is already organized that way. A desk has a mandate (profile), a blotter and research archive (memory), a morning meeting (planning), and an order-management system (action). A financial-crime unit has a risk appetite, a case system, a typology library, and filing channels. Treasury has a liquidity policy, a cash ledger, a forecast book, and payment rails. The agent literature is, in that sense, rediscovering the firm.

Two cautions follow. First, language-agent evaluations outside finance often use games, shopping sites, or question-answering sandboxes. Those environments forgive errors that a live book will not. Second, once several agents talk to one another, the system inherits the classical pathologies of multi-agent design (collusion, herding, conflicting objectives, and responsibility diffusion) on top of the newer pathologies of hallucination, prompt injection, and tool-call misfires. Even research on generative agents as interactive simulacra of human behavior, which is not a finance paper, is a reminder that multi-agent conversation can look coherent while drifting from any single mandate³⁸.

2.2 Agentic systems in trading and investment

Equity and futures research has moved fastest, because simulated trading is easy to score. Lopez-Lira and Tang documented that a general-purpose chat model can classify the return implication of news headlines and that the resulting long–short signal was informative in their sample⁶. That paper is about *perception*, not agency: the model does not send the order. The agent papers close that loop.

FinMem equips an LLM trader with a layered memory and a character design so that short-horizon ticks, mid-horizon news, and long-horizon theses are not dumped into a single context window⁷. FinAgent, presented as a multimodal, tool-augmented foundation agent for trading, consumes text and visual market information and reports, on one of its datasets, a 92.27 percent return and a large relative improvement versus

the authors' baselines⁸. TradingAgents instead copies the *firm*: fundamental, sentiment, and technical analysts, plus traders with different risk profiles, debate before a decision⁹. Later work decomposes investment analysis into finer tasks rather than coarse role prompts, on the argument that abstract “be a portfolio manager” instructions hide the actual workflow.

A 2026 survey of LLM trading agents, framed as an audit-oriented evidence map, is blunt about the state of play¹¹. A parallel finance-wide survey of agentic systems reaches a similar conclusion at the industry level: architecture is outrunning evaluation and regulatory design¹². Reviews of LLM agents in banking likewise organize applications into a small number of operational clusters rather than a single “AI trading” story³⁷. Architectural variety has grown quickly. Comparable evaluation protocols, execution semantics, and reproducible artifacts have not. That diagnosis is the right one for any reader tempted to treat a single backtest as a license to automate a book. Simulated fills, ignored market impact, look-ahead in news timestamps, and non-replicable prompts can manufacture agency that would not survive a live, capacity-constrained book.

2.3 Fraud, financial crime, and investigative work

Fraud and anti-money-laundering (AML) work is already a human–machine loop. Production stacks combine rules, unsupervised anomaly detection, supervised models, network features, and investigator judgment. A 2024 BioCatch industry survey reported that 73 percent of organizations were using AI for fraud detection and 74 percent for financial-crime detection, with machine learning, natural language processing, and deep learning all in wide use¹⁶. Those figures describe *model* adoption, not agentic case handling.

The agentic turn in this domain is investigative. Systems that compile entity research, draft narratives, gather open-source and consortium context, and walk a standard enhanced-due-diligence checklist look less like a transaction-scoring model and more like a junior analyst. Axelsen and coauthors describe an agentic architecture for financial-crime compliance that aims at end-to-end orchestration, including case management, rather than a single detection score¹⁷. Industry write-ups of investigative copilots make the same point at a lower formal level: the scarce resource is investigator time, and the first safe automation is assembly of the file, not the suspicious-activity decision.

Two structural facts make this domain different from trading. False positives are expensive in staff time; false negatives are expensive in loss, regulatory censure, and, in the limit, charter risk. And many downstream actions (blocking a payment, exiting a customer, filing a report) are only partly reversible and are legally attributed to the institution, not to a model vendor.

2.4 Cash-flow forecasting and treasury

Treasurers have said for years that cash forecasting is among their hardest routine problems. The inputs are heterogeneous (enterprise-resource-planning invoices, bank statements, cards, market-margin calls, tax calendars), the horizon mix is awkward (intraday liquidity versus monthly planning), and the loss function is asymmetric (a missed funding gap is not the same as idle cash). Industry notes from banks and treasury platforms describe machine-learning ensembles as already in use for pattern recognition on those series. J.P. Morgan's treasury commentary is representative: neural nets, random forests, and related models are presented as reducing forecast error relative to purely statistical baselines in corporate case material¹⁸.

What is *not* yet dense is a literature on *agentic* treasury. A 2025 review of generative AI in cash-flow forecasting concluded that empirical research remains scarce relative to the much larger earnings-forecasting literature¹⁹. That gap matters. A forecasting model that writes a number into a dashboard is an R0/R1 object. An agent that reconciles bank files, explains a variance, proposes a commercial-paper issue, and,

if permitted, instructs a payment bank is an R2 object sitting on the institution’s liquidity mandate. The second object needs a different control design even if it uses the first object as a tool.

2.5 Risk early-warning

Early-warning systems (EWS) are the oldest “act before default” technology in banking. Classical versions used financial ratios, expert watch-lists, and later discrete-choice models. Machine learning entered at two scales. At the *system* scale, central-bank and supervisory papers train classifiers on confidential and public indicators to predict bank distress; examples include Bank of England staff work on the United Kingdom²⁰ and, more recently, European Banking Authority staff work on large European banks using neural nets, trees, and forests under a carefully defined distress label²¹. At the *customer* scale, banks monitor commercial and retail books for pre-default deterioration; recent studies argue for data-driven EWS rather than purely judgmental watch-lists²⁵. European Central Bank work has also treated early-warning *modeling* itself as a design problem (alert windows, preference between missed crises and false alarms) rather than a single AUC contest^{22,23}.

Industry notes describe the same movement toward richer, more timely EWS²⁴. Generative and agentic layers are being discussed on top of that stack (summarizing why a name deteriorated, gathering news and covenant evidence, drafting a credit-committee note), but the statistically identified object remains a risk score or a hazard. Autonomy here collides directly with consumer-credit regulation, because an agent that converts an early-warning signal into a limit cut or a declined renewal is making a creditworthiness decision about a person or a firm.

2.6 Customer interaction

Conversational systems in retail banking began as menu trees and frequently-asked-question bots. Systematic reviews of financial chatbots document efficiency and availability gains together with persistent problems of trust, escalation, and handling of complex or distressed clients. LLM agents change the surface: they can maintain multi-turn state, call account tools, and speak in a more natural register. They also create a new failure mode. An industry-track study of customer-support LLM agents treats adherence to standard operating procedures as a first-class evaluation target, distinguishing a static prompt agent from agents that are more tightly bound to procedures³⁴. That is the right instinct for a regulated conversation. A service agent that is merely “helpful” can waive a fee it should not waive, give investment-like guidance, or mishandle a vulnerable customer.

The research and vendor records also show that high-volume, low-risk fulfillment (card controls, balance questions, document collection) is where autonomy is most plausible, and that complaints, collections, advice, and financial-difficulty conversations are where human ownership should remain visible.

2.7 The gap this paper fills

Existing papers optimize a domain. Trading papers maximize a backtest statistic. Crime papers orchestrate a case. Treasury notes reduce forecast error. EWS papers trade off false alarms against missed distress. Service papers measure containment or procedure adherence. Supervisory texts define model risk and, in Europe, high-risk AI. What is missing is a *portable* account of authority that a chief risk officer, a desk head, and a researcher can share. The next section states that account.

3. A decision-rights framework

3.1 What counts as an agent here

For this paper, a financial agent is a closed-loop policy that, at each decision time t , maps an observation o_t , a memory m_t , and a written mandate M into an action a_t , a justification j_t , and an escalation flag e_t :

$$\pi: (o_t, m_t, M) \mapsto (a_t, j_t, e_t). \quad (1)$$

A *predictive model* is the special case in which a_t is only a score or a label and some other system acts. A *copilot* is the special case in which a_t is a draft and e_t is always on. A *closed-loop agent* is allowed, inside M , to change a book, a case, a forecast that others will use, or a promise to a client. A *multi-agent system* is a set of such policies plus a protocol for who speaks, who may overrule whom, and how disagreement is recorded.

This definition is deliberately institutional. The mandate M is not a prompt. It is the intersection of product policy, risk appetite, license, and system entitlements. If the mandate is only in the prompt, the institution does not have an agent; it has an unbounded script.

3.2 The agency spectrum

Figure 1 places systems on a spectrum that trading surveys have begun to use and that generalizes cleanly.

1. **Prediction models** perceive and score (a default probability, a fraud score, a sentiment class).
2. **Signal generators** add a recommended action but do not touch the book.
3. **Partial agents** plan and remember but stop before irreversible action.
4. **Closed-loop agents** complete perceive–plan–act–evaluate under limits.
5. **Multi-agent ecosystems** add specialization and bargaining across roles or firms.

Moving right raises option value and supervisory burden together. The interesting design region for the next few years is not the far right. It is the disciplined occupation of regions 3 and 4.

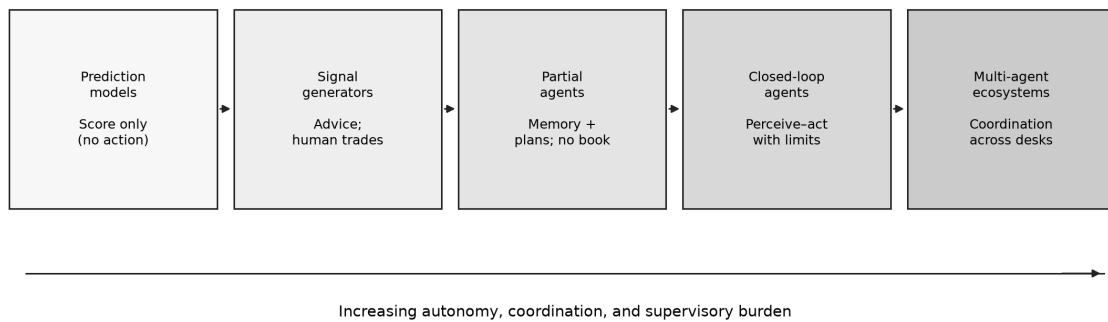


Figure 1: Conceptual. Agency spectrum from scoring models to multi-agent ecosystems. Moving right raises option value and supervisory burden together. This is a design schematic, not an estimated index.

3.3 Decision rights R0–R4

The operational content of the spectrum is a rights ladder.

- **R0 (Perceive).** Read data, update features, write a score or a summary. No outward commitment.
- **R1 (Recommend).** Propose an action, a case disposition, a forecast, or a client reply, with a justification, to a human or to a rules gate.
- **R2 (Execute, bounded).** Take an action inside a pre-registered action set $\mathcal{A}(M)$, a notional or severity cap, a reversibility class, and a working-hours or venue constraint.
- **R3 (Update policy).** Change thresholds, prompts, tools, memory-writing rules, or the agent's own objective weights. This is model change, not business-as-usual execution.
- **R4 (Coordinate).** Delegate to, debate with, or overrule other agents, including agents in another legal entity.

R0 and R1 are the default industrial setting. R2 is the first true autonomy and should be rare, small, and reversible unless a human is still on the hook in real time. R3 should pass through the same change-management path as a production model. R4 should be treated as the creation of a new organizational unit, not as a clever prompt pattern.

Figure 3 applies this ladder to the five domains. The figure is a *design map*, not a measured maturity score.

3.4 The control loop

Figure 2 is the reference architecture. Four operational stages sit above a mandate gate; an audit layer and an accountability layer sit below.

- **Sense.** Ingest market data, ledger events, alerts, documents, and client utterances. Normalize them into objects the rest of the loop can cite.
- **Plan.** Retrieve memory and policy, reason, simulate if a sandbox exists, and, where designed, run a structured debate among specialist agents.
- **Act.** Emit a tool call that is either a recommendation, a bounded execution, or an escalation.
- **Learn.** Write an episode log, score the outcome against a delayed label, and update *memory*. Updating *policy* is not learning in this sense; it is R3. Verbal self-reflection loops of the Reflexion type belong here as memory writes unless they change the approved policy, in which case they are R3¹⁰.

The **mandate gate** is the object most vendor diagrams omit. It checks three things before any R2 action: that the proposed a_t lies in $\mathcal{A}(M)$; that a scalar supervisory statistic s_t (uncertainty, novelty, limit utilization, client vulnerability, or a blended score) is below a threshold s^* ; and that no kill-switch or dual-control lock is active. Formally, execution authority is

$$\alpha_t = \mathbf{1}\{a_t \in \mathcal{A}(M)\} \cdot \mathbf{1}\{s_t < s^*\} \cdot \mathbf{1}\{\kappa_t = 0\}, \quad (2)$$

where $\kappa_t = 1$ if a kill switch, freeze, or four-eyes requirement is on. If $\alpha_t = 0$, the system may only recommend or escalate.

3.5 An institutional loss function

Autonomy should be sized against a loss that the institution already understands. A useful sketch is

$$\mathbb{E}[L] = \lambda_F \mathbb{E}[L^{\text{false}}] + \lambda_M \mathbb{E}[L^{\text{miss}}] + \lambda_\tau \mathbb{E}[\tau] + \lambda_H \mathbb{E}[c^{\text{human}}] + \lambda_S \mathbb{E}[L^{\text{sys}}], \quad (3)$$

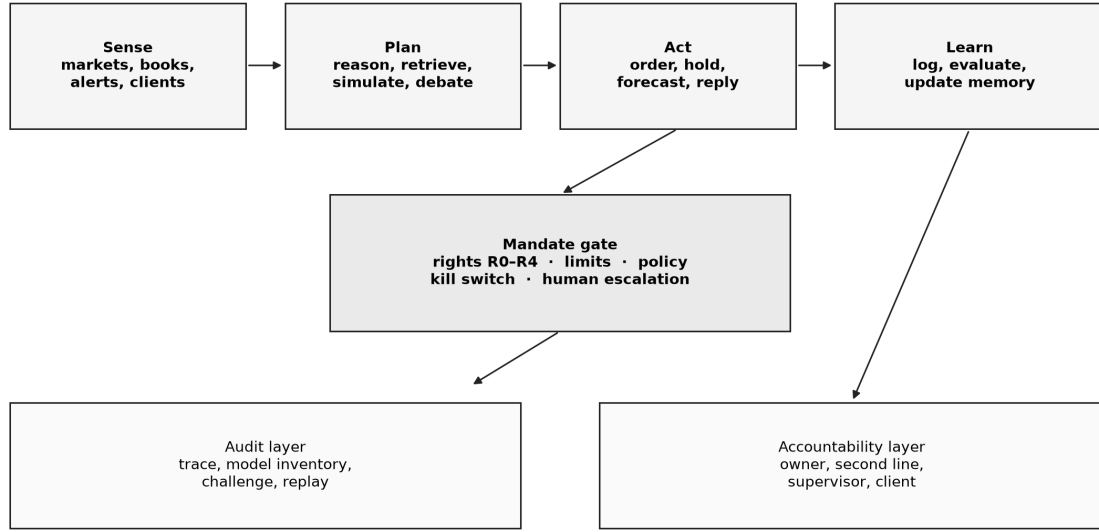


Figure 2: Conceptual. Reference control loop. Operational stages sit above a mandate gate. Audit and accountability layers make replay and named ownership possible. Not an estimated process model.

where L^{false} is the cost of acting in error (a bad trade, a blocked good customer, a wrong promise), L^{miss} is the cost of not acting (fraud loss, missed distress, unanswered client, unhedged gap), τ is decision latency, c^{human} is the cost of the residual human process, and L^{sys} is a systemic or franchise term (herding, correlated model failure, conduct fine, operational outage). The λ weights are not statistical hyperparameters. They are risk-appetite choices, and they differ by domain. Trading usually loads λ_F and λ_S . Fraud and early-warning load λ_M . Service loads λ_F when the error is a bad promise and λ_τ when the error is abandonment. Treasury loads both gap risk and the opportunity cost of idle cash.

Two implications are immediate. First, raising autonomy to cut c^{human} is rational only if it does not increase L^{false} and L^{sys} by more. Second, evaluation that reports only hit rate, Sharpe, or containment is incomplete. It omits the terms that make finance different from a demo.

3.6 Human-agent residual responsibility

Even at R2, residual responsibility cannot be deleted by a disclaimer in a model card. Someone named must own the mandate, the threshold s^* , the action set, and the after-action review. The justification j_t is not decoration. It is the object audit, challenge, and, in consumer settings, explanation duties will inspect. If j_t is an unfaithful chain of thought, the institution has a documentation failure, not a clever agent.

4. Mapping the framework to five financial functions

This section applies the ladder. Table 1 summarizes the design map. The subsections state what can be automated now, what should remain gated, and what would count as a reckless grant of rights.

4.1 Trading

Trading is the most visible agentic research domain and the easiest place to over-learn from a backtest. Perception (R0) is already industrial: signals, alternative data, and, after Lopez-Lira and Tang, news interpretation by general-purpose models. Recommendation (R1) is the natural home of FinMem-style memory agents and of specialist analysts in a TradingAgents-like debate: they write a note, a target, and a dissent. Bounded execution (R2) is old in another costume. Electronic market-making and execution

	R0 Perceive	R1 Recommend	R2 Execute (bounded)	R3 Update policy	R4 Coordinate
Trading	Default target	Default target	Default target	Conditional with gates	Conditional with gates
Fraud / AML	Default target	Default target	Conditional with gates	Conditional with gates	Conditional with gates
Cash-flow forecasting	Default target	Default target	Conditional with gates	Conditional with gates	Conditional with gates
Risk early-warning	Default target	Default target	Conditional with gates	Rare / human-only	Conditional with gates
Customer interaction	Default target	Default target	Conditional with gates	Rare / human-only	Conditional with gates

Figure is a design map, not a measured capability score.

Figure 3: Conceptual. Decision-rights design map across five functions. Shading is prescriptive guidance, not a measured capability or maturity score.

algorithms have long sent live orders under limits, venues, and kill switches. What is new is an LLM in the planner, not the existence of automatic orders.

The safe R2 envelope for a language agent is therefore narrower than a fully engineered execution algorithm. It should look like this: the agent may choose among pre-approved tactics (join, take, stay flat, reduce) inside a name list, a gross and net notional, a maximum participation rate, and a maximum daily loss. It may not widen the universe, raise its own limits, or turn itself on after a kill switch. Policy updating (R3), meaning a change to the prompt, the debate protocol, or the risk character, is a model release. Multi-agent coordination (R4) across a simulated desk is a research device; across live, separately mandated books it is a correlated-risk event and should be inventoried as such.

Evaluation should follow the 2026 trading-agent surveys, not social-media Sharpe ratios: stated execution assumptions, no look-ahead, capacity, transaction-cost models, seed and prompt logs, and a hold-out *regime*, not only a hold-out ticker list. Author-reported gains, including FinAgent's large improvement on one dataset, are encouragement to keep studying the architecture. They are not a production mandate.

4.2 Fraud detection and financial-crime compliance

Here the first high-value agent is a file assembler, not a judge. R0 is the existing scoring and graph stack. R1 is a copilot that retrieves prior cases, beneficial-ownership fragments, typologies, and a draft narrative with citations to the systems of record (the direction described in investigative-copilot deployments and in agentic compliance research). R2 is appropriate for a small, reversible set: additional authentication, a temporary hold under already approved rules, or auto-closure of alerts that fail a conservative, documented test. R2 is not appropriate for customer exit, suspicious-activity filing, or sanctions disposition without a named officer.

The BioCatch survey numbers show that AI *detection* is already mainstream. The remaining bottleneck is the human investigation factory. Agents help that factory if they reduce the time to a *well-sourced* decision. They harm it if they produce fluent but ungrounded narratives that investigators rubber-stamp. Hence the audit layer in Figure 2 must store the retrieved passages, not only the final paragraph. R3 (retraining a detector, changing an auto-close rule, editing a typology library) is model and policy change. R4 (agents

that query consortium data or other institutions) creates data-sharing and liability questions that sit above the model.

Note the legal asymmetry created by the EU AI Act: systems used *to detect financial fraud* are carved out of the Annex III credit-scoring high-risk item. That carve-out is not a holiday from AML, privacy, or operational-resilience law, and it does not cover an agent that uses the same stack to cut a person's credit.

4.3 Cash-flow forecasting

A modern treasury forecast is already a pipeline: ingest bank and enterprise data, classify flows, estimate timing, apply overlays, publish a liquidity view, and explain variances. Machine-learning models belong at R0. An agent that reconciles breaks, writes the variance commentary, and proposes a funding or investment action belongs at R1. R2 should be limited to mechanical, low-stakes acts that the treasury policy already allows a junior officer to perform without a meeting. Examples include sweeping cash among prepaid accounts under a documented waterfall, or generating (but not sending) payment files for a four-eyes release.

Two reasons counsel against a bold R2. First, the scholarly base for *generative* cash-flow agents is still thin relative to earnings research, so institutions are generalizing from vendor case studies more than from cumulative science. Second, a wrong forecast that is merely displayed is a planning error; a wrong forecast that automatically draws a facility or fails to pre-fund a margin call is a liquidity event. R3 covers recalibration of the forecast model and any change to the agent's overlay rules after a structural break (a new enterprise-resource-planning system, a merger, a tax reform). R4 appears when the treasury agent starts instructing trading, payables, or subsidiary agents; that is a group-liquidity protocol and needs a group-liquidity owner.

4.4 Risk early-warning

EWS work is statistically mature and legally sensitive. R0 is the score: borrower-level deterioration, name-level early warning, or, in supervisory research, bank-distress probability. R1 is the packet: why the score moved, which covenants or news items fired, what comparable names did, and what the recommended watch-list or limit action is. That packet is a natural language-agent task sitting on top of the older classifiers used in Bank of England, European Banking Authority, and bank-internal studies.

R2 should be confined to *internal risk hygiene* that does not change the customer's legal position: placing a name on a watch-list, requesting information, or raising a monitoring frequency. Changing a limit, calling a loan, re-pricing, or declining a renewal is a credit decision. For natural persons in the EU it is presumptively a high-risk AI use. Even outside that statute, it is a decision that credit committees exist to own. R3 is recalibration of the EWS, including the alert-window and preference parameters that ECB-style early-warning design treats as first-class. R4 is dangerous in this domain because correlated early-warning agents can produce simultaneous limit cuts, a textbook procyclicality channel.

The loss-function weights here are not symmetric. A tradition of supervisory EWS work treats missed distress as the event one is willing to spend false alarms to avoid. An agent optimized for investigator convenience will invert that preference unless the mandate says otherwise.

4.5 Customer interaction

Customer-facing agents are where autonomy is most visible and most easy to over-extend. R0 is intent detection and retrieval from a knowledge base. R1 is a draft reply or a recommended next step for a human banker. R2 is appropriate for authenticated, low-ambiguity fulfillment: card on/off, copy of a statement, status of a payment, collection of a missing document. It is not appropriate for advice that could be read as a personal recommendation, for collections strategy, for hardship and vulnerability, or for a complaint that may become a conduct file.

Procedure adherence, as recent industry evaluations emphasize, is the binding quality metric, not conversational charm. The mandate must include a product map (what the bot is allowed to know), an identity and privacy design (what the language model is allowed to see), and a hard escalation ontology (when e_t must flip). R3 includes changing the dialogue policy, the knowledge articles, or the tool entitlements after a product launch. R4 appears in “agent swarms” that hand a client from service to fraud to collections without a single human owner. That pattern recreates the worst of interactive-voice-response systems unless a named accountable person remains on the episode.

5. Architecture, evaluation, and failure modes

5.1 A common stack

Despite domain differences, production designs converge.

- **Objects, not prompts, as the system of record.** Trades, alerts, forecasts, names, and conversations must live in ledgers the agent can cite and the auditor can replay. If the only memory is the context window, the institution cannot reconstruct a decision.
- **Tools as entitlements.** Every application programming interface the agent can call is an access-control question. Read tools implement R0. Write tools implement R2 and must be wrap-able by α_t in (2).
- **Specialists over generalists.** The trading literature’s move from one “trader” prompt to analyst/trader/risk roles, and the crime literature’s move from one detector to investigator/writer/checker roles, is the same idea: decompose so that disagreement is visible.
- **An evaluator that is not the actor.** A second model or a rules pack that scores j_t for grounding, policy breach, and toxicity should not share a gradient, a prompt, or an incentive with the actor. This is the internal analog of independent model validation.
- **Time as a first-class input.** Markets, fraud rings, cash calendars, and credit cycles are non-stationary. Memory that cannot forget, and policies that cannot be dated, will overfit last month’s regime.

5.2 What to measure

Table 2 lists evaluation objects that match the loss function in Section 3. The important methodological point is that *decision quality* is not *language quality*. A fluent justification of a bad action is a worse institutional outcome than an awkward escalation. Trading surveys that complain about non-comparable protocols are making this point for one domain; it applies to all five.

Replay is the shared gold standard. If the institution cannot freeze the data, the tools, the mandate version, and the random seeds and obtain the same triple (a_t, j_t, e_t) from (1), it cannot validate, it cannot explain, and it cannot run a fair after-action review. That is a higher bar than storing chat logs.

5.3 Characteristic failures

Several failures are already predictable from the combined literatures.

- **Unfaithful reasons.** Chain-of-thought text need not be the causal path to the tool call. Treating j_t as an explanation without intervention tests is a governance error.
- **Tool-call smuggling and prompt injection.** Any agent that reads email, news, or client free text can be instructed, by that text, to call a write tool. The mandate gate must assume hostile content.
- **Silent entitlement creep.** A helper tool added for “just this investigation” becomes a production write path. R3 discipline exists to stop that.

- **Herdning.** If many desks or many banks use similar agents on similar public news, Lopez-Lira-and-Tang-style predictability decays and impact concentrates. That is a market-structure issue, not only a model issue.
- **Procyclical early warning.** Agents that all raise watch-list intensity in the same week create the appearance of omniscience and the reality of a credit crunch.
- **Responsibility diffusion.** Multi-agent debate can be an excuse (“the risk agent signed off”) unless one named book-runner remains liable.
- **Demo-to-production leakage.** Author-reported returns, vendor cycle-time claims, and executive surveys are not substitutes for a validation pack.

6. Governance and regulation

6.1 Agents are models, and more than models

U.S. banking supervisors have long defined a model as a quantitative method that processes inputs into estimates used in decisions. An LLM agent that plans and writes is an awkward fit for that sentence, but the *use* is not: the institution is still relying on a statistical engine to change a decision. SR 11-7’s core demands (inventory, conceptual soundness, ongoing monitoring, outcomes analysis, and independent challenge) therefore attach²⁹. The 2026 interagency update, issued after more than a decade of supervisory experience and after a wave of machine-learning and generative systems, is best read as a reminder that those principles are risk-based and technology-agnostic, not as a claim that agents are outside model risk³⁰.

What agents add, and what older model-risk packs under-weight, is *tool entitlement*, *prompt and memory versioning*, *non-determinism*, and *cross-agent protocol*. Those belong in the model inventory as first-class fields, next to training data and performance tests.

6.2 European high-risk AI, operational resilience, and privacy

Regulation (EU) 2024/1689 (the AI Act) entered into force in August 2024²⁶. Annex III lists, among high-risk systems, AI used to evaluate the creditworthiness of natural persons or to establish their credit score, and carves out systems used to detect financial fraud^{26,35,36}. The European Banking Authority has begun mapping those high-risk duties onto existing sectoral rules for creditworthiness²⁷. For this paper’s five domains the practical reading is:

- A risk-early-warning agent that changes a consumer’s access to credit is in the high-risk conversation.
- A fraud agent that only detects and investigates is in a different, though still regulated, bucket.
- A customer agent that profiles a person for credit or insurance pricing is not “just a chatbot.”
- Transparency, logging, human oversight, and data-governance duties will be read together with the General Data Protection Regulation’s limits on solely automated decisions that produce legal or similarly significant effects³³. NIST’s AI risk-management framework and ISO/IEC 42001 offer complementary, non-statutory management-system language that many groups already map onto model risk^{31,32}.

The Digital Operational Resilience Act (DORA) adds a second European pillar: an agent that can instruct payments, trading venues, or case systems is an information and communication technology risk, not only an ethics risk²⁸. Concentration on a single model vendor, a single cloud, or a single prompt-orchestration platform is an outsourcing and continuity issue.

6.3 Conduct, markets, and crime

Algorithmic-trading and market-abuse regimes already expect testing, kill switches, and responsibility for machine-generated orders⁴⁰. Putting a language model in the planner does not relax those expectations. AML and sanctions regimes expect a documented, risk-based program and a human officer³⁹; an agent can prepare the file. Consumer-conduct regimes expect fair treatment and careful handling of vulnerability; a charming agent that hurries a distressed customer is a conduct event.

Figure 4 stacks these constraints. External accountability sits on institutional control, which sits on runtime control, which sits on the agent stack, which sits on systems of record. Skipping a layer is how a prototype becomes an incident.

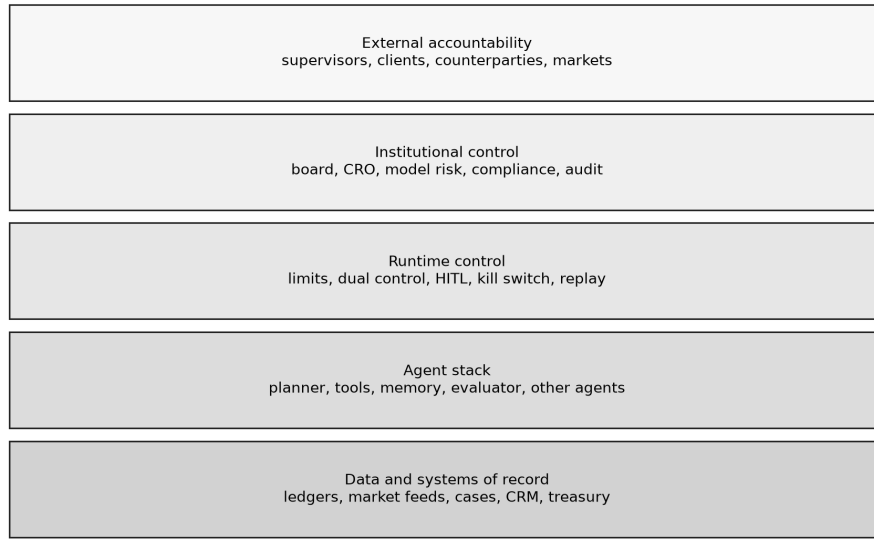


Figure 4: Conceptual. Five-layer governance stack from systems of record to external accountability. Skipping a layer is how a prototype becomes an incident. Not an estimated organizational model.

6.4 Organization

Three organizational facts decide whether the framework is real.

1. **A named owner per mandate**, not per model vendor. The desk head owns the trading agent's $\mathcal{A}(M)$. The money-laundering reporting officer owns filing recommendations. The treasurer owns cash movements. The chief credit officer owns limit-changing early warning. The head of service owns promises to clients.
2. **Second-line rights over s^* and $\mathcal{A}(M)$** . If only the first line can widen the action set, autonomy will ratchet.
3. **A single inventory**. Agents, prompts, tools, and evaluation sets should not live in a shadow information-technology register separate from the model-risk inventory.

These are unglamorous. They are also the difference between a research program and an uncontrolled automated officer.

7. Research agenda

The synthesis points to work that would change practice, not only add another named architecture.

1. **Shared execution semantics.** Trading-agent surveys have already called for this. The same demand should be made of crime, treasury, EWS, and service agents: state what the action *does* in the world, what it costs to reverse, and how fills, holds, or promises are simulated.
2. **Decision-rights experiments, not only accuracy experiments.** A paper that compares R1 versus R2 under a fixed loss $\mathbb{E}[L]$ from (3) is more informative for a chief risk officer than a paper that reports a new headline return.
3. **Faithfulness tests for j_t .** Interventions on retrieved evidence, counterfactual tool masking, and human-judge protocols are needed before justifications are treated as explanations.
4. **Systemic and many-agent evaluation.** Herding in news-driven trading, correlated limit cuts in EWS, and common-mode vendor failure belong in test packs, including agent-based market models.
5. **Treasury and EWS agent benchmarks.** These two domains are operationally central and scientifically thin at the *agent* layer. Public, non-proprietary cash and (suitably anonymized) early-warning challenge sets would correct a research portfolio that is overweight equities.
6. **Procedure-bound service agents.** Containment rate without procedure breach, privacy leakage, and escalation quality should replace “human-like” as the lead metric.
7. **Model-risk methods for non-deterministic, tool-using systems.** Challenger regimes, ongoing monitoring, and conceptual-soundness review need concrete methods when the “model” is a prompt plus tools plus memory.
8. **Human residual skill.** If agents take the routine file, institutions must measure whether humans still catch the non-routine miss (the analogue of automation complacency in other safety-critical fields).

8. Conclusion

Agentic artificial intelligence will not enter finance as a single omniscient officer. It will enter as a set of specialized loops that read a book, a case, a forecast, a watch-list, or a conversation and try to do the next step. Research prototypes in trading, investigative copilots in financial crime, machine-learning cash forecasts, early-warning classifiers, and procedure-bound service agents are five faces of that same object.

The mistake is to treat the object as a better chatbot or a better signal. It is a claimant on decision rights. This paper’s framework makes those rights explicit. Perceive and recommend widely. Execute only inside a registered action set, a supervisory statistic, and a kill switch, as in (2). Change policy through model governance. Allow agents to command other agents only when the institution is prepared to inventory a new organizational unit. Size every grant of rights against the loss in (3), which includes error, delay, human cost, and systemic residue, not against a demo metric.

Executives already know, in survey form, that they want the productivity and do not yet trust the transaction. Supervisors already have, in model-risk guidance, the AI Act, operational-resilience law, and conduct rules, most of the hooks they need. The remaining work is design: put the mandate in the gate, put the evidence in the log, and put a name on the authority. Institutions that do that can use agents. Institutions that skip it will discover that autonomy without rights is just speed in the wrong direction.

References

- [1] L. Wang et al., “A survey on large language model based autonomous agents,” arXiv:2308.11432, 2023.

Table 1: Conceptual. Decision-rights map by domain. Design guidance only. Not an estimated capability score.

Domain	Default R0/R1	Acceptable R2	Generally reserved
Trading	Signals, debate notes, target prices	Tactics inside name list, notional, loss, and participation caps	Universe expansion; self-raised limits; live R3
Fraud / AML	Scores, entity research, draft narratives	Holds and auto-close under conservative, documented rules	Filing, exit, sanctions disposition
Cash-flow	Ingest, classify, forecast, explain	Policy-bound sweeps; payment <i>preparation</i>	Unattended drawing of facilities
Risk EWS	Scores, watch packets, committee drafts	Internal watch-list and monitoring frequency	Limit cuts or declines that change a customer's position
Customer	Intent, retrieval, draft replies	Authenticated, low-ambiguity fulfillment	Advice, hardship, complaints, collections strategy

Table 2: Conceptual. Evaluation objects aligned to the institutional loss in (3). Cells are loss types, not measured outcomes of this paper.

Object	Trading	Fraud / AML	Cash-flow	EWS	Customer
Error of commission	Bad fill, unintended inventory	Wrong hold, wrong close	Idle-cash or wrong-sweep cost	Unnecessary tightening	Wrong promise or waiver
Error of omission	Missed hedge or alpha (secondary)	Missed typology, late file	Missed gap, missed call	Missed distress	Abandoned or silent failure
Latency	Slippage versus arrival	Time-to-decision on true hits	Intraday update lag	Alert-to-committee time	Time-to-resolution
Human residual	Override quality	Investigator disagreement	Treasurer overlay quality	Committee dissent	Escalation quality
Systemic / franchise	Herding, impact	Consistency of typology	Group-liquidity spiral	Procyclical limit cuts	Conduct and privacy events
Replay	Prompts, fills, costs	Evidence graph	Bank-file vintage	Score vintage	Tool entitlements, SOP version

- [2] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, “ReAct: Synergizing reasoning and acting in language models,” in *Proc. ICLR*, 2023.
- [3] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2020.
- [4] M. Wooldridge, *An Introduction to MultiAgent Systems*, 2nd ed. Wiley, 2009.
- [5] J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” in *Proc. NeurIPS*, 2022.
- [6] A. Lopez-Lira and Y. Tang, “Can ChatGPT forecast stock price movements? Return predictability and large language models,” arXiv:2304.07619, 2023.
- [7] Y. Yu et al., “FinMem: A performance-enhanced LLM trading agent with layered memory and character design,” arXiv:2311.13743, 2023.
- [8] W. Zhang et al., “A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist,” arXiv:2402.18485, 2024.
- [9] Y. Xiao, E. Sun, D. Luo, and W. Wang, “TradingAgents: Multi-agents LLM financial trading framework,” arXiv:2412.20138, 2024.
- [10] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: Language agents with verbal reinforcement learning,” arXiv:2303.11366, 2023.
- [11] Y. Xia, P. You, T. Wang, F. Liu, H. Qi, X. Wu, and S. Zhang, “Agentic trading: When LLM agents meet financial markets,” arXiv:2605.19337, 2026.
- [12] I. Aldridge et al., “Agentic artificial intelligence in finance: A comprehensive survey,” arXiv:2604.21672, 2026.
- [13] PwC, “PwC’s AI agent survey,” May 2025.
- [14] Capgemini Research Institute, “Rise of agentic AI: How trust is the key to human–AI collaboration,” 2025.
- [15] S&P Global, “Beyond automation: Agentic AI and scaling fragmented financial markets,” 2025.
- [16] BioCatch, “2024 AI fraud and financial crime survey,” 2024.
- [17] H. Axelsen et al., “Agentic AI for financial crime compliance,” arXiv:2509.13137, 2025.
- [18] J.P. Morgan, “AI-driven cash flow forecasting: The future of treasury,” 2024.
- [19] Z. Wahyudi, “Generative AI and the future of cash flow forecasting,” *Escalate*, 2025.
- [20] J. Suss and H. Treitel, “Predicting bank distress in the UK with machine learning,” Bank of England Staff Working Paper, 2019.
- [21] D. Malikkidou and W. Strohbach, “Predicting bank distress in Europe using machine learning and a novel definition of distress,” EBA Staff Paper Series no. 21, Jan. 2025.
- [22] J. Lang, T. Peltonen, and P. Sarlin, “A framework for early-warning modeling with an application to banks,” ECB Working Paper, 2018.

- [23] A. Petropoulos, V. Siakoulis, E. Stavroulakis, and N.E. Vlachogiannakis, “Predicting bank insolvencies using machine learning techniques,” *International Journal of Forecasting*, vol. 36, no. 3, pp. 1092–1113, 2020.
- [24] EY, “The future of early warning systems in banking,” Feb. 2025.
- [25] J. Beltman et al., “Predicting retail customers’ distress in the finance industry,” *Journal of Retailing and Consumer Services*, 2025.
- [26] Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act), 2024.
- [27] European Banking Authority, “AI Act: implications for the EU banking and payments sector,” 2025.
- [28] Regulation (EU) 2022/2554 (Digital Operational Resilience Act).
- [29] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, “Supervisory guidance on model risk management” (SR 11-7 / OCC 2011-12), Apr. 2011.
- [30] Office of the Comptroller of the Currency, “Model risk management: Revised guidance,” OCC Bulletin 2026-13, Apr. 2026.
- [31] NIST, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” 2023.
- [32] ISO/IEC 42001:2023, *Information technology – Artificial intelligence – Management system*.
- [33] Regulation (EU) 2016/679 (General Data Protection Regulation), Art. 22.
- [34] S. Balaji et al., “Beyond IVR: Benchmarking customer support LLM agents,” in *Proc. EACL Industry*, 2026.
- [35] G. Mazzini and F. Bagni, “Considerations on the regulation of AI systems in the financial sector by the AI Act,” *Frontiers in Artificial Intelligence*, vol. 6, 2023.
- [36] P. Hacker, “The future of credit underwriting and insurance under the EU AI Act,” *Harvard Data Science Review*, 2025.
- [37] D. Batra, “A review of LLM agent applications in finance and banking,” SSRN working paper, 2025.
- [38] J.S. Park et al., “Generative agents: Interactive simulacra of human behavior,” in *Proc. UIST*, 2023.
- [39] Financial Action Task Force, *International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation* (the FATF Recommendations), updated editions.
- [40] Directive 2014/65/EU (MiFID II), provisions on algorithmic trading.