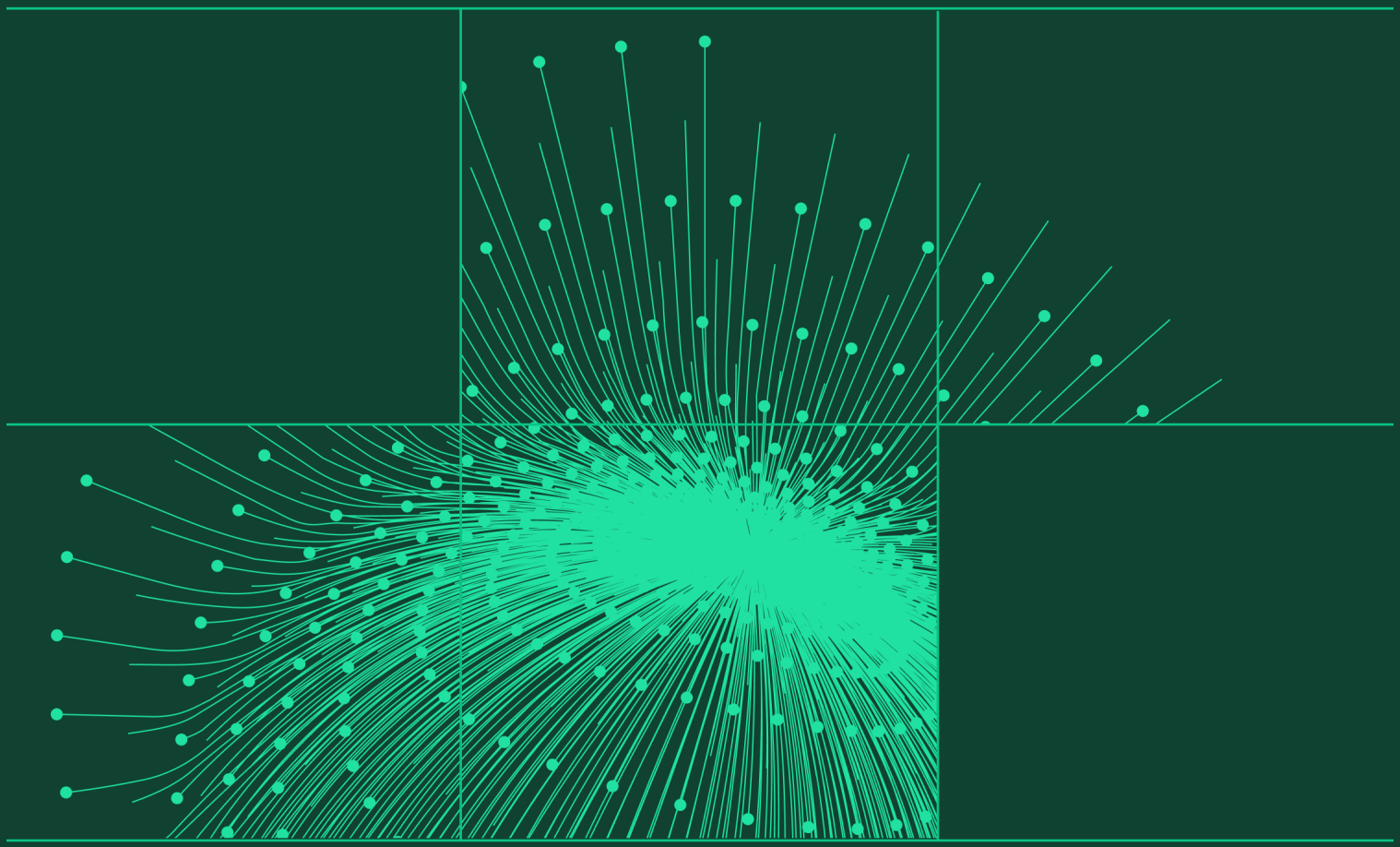
WHITEPAPER

Deploying Agentic AI in Financial Services

Reducing regulatory risk and ensuring audit readiness

VOLUME 2

Sardine



Deploying Agentic AI in Financial Services

Reducing regulatory risk and ensuring audit readiness

Authors:
Simon Taylor, Soups Ranjan, PhD, Ravi Loganathan, Matt Vega (3CI, SIP, ECFE, CPFPP), Ryan McCormack, and Erich Reich.

Reviewers:
David Silverman

Volume 2

Table of Contents

Abstract 5
1. The Agentic Oversight Framework Introduction 7
2. The Benefits of AI Agents for Compliance 9
3. How does an AI Agent differ from other “Models”? 10
4. The Compliance Challenge 14
5. The Agentic Oversight Framework (AOF) in depth 17
6. Reducing Regulatory Risk and Ensuring Audit Readiness 25
7. Use Cases and Case Studies of the AOF in Action 29
   KYC Onboarding 30
   Sanctions and Pep Alerts 34
8. Further Development and Collaboration 34
9. About the Authors 35
References 36

This paper presents findings from Sardine's deployment of AI agents in live production environments across multiple financial institutions, where they operated for over three months, in BSA/AML compliance workflows. AI agents, or agentic AI, refer to artificial intelligence systems that can reason, adapt to context, and take goal-directed actions with minimal human intervention. We observed that these agents can significantly speed up manual reviews for Know Your Customer (KYC) onboarding and sanctions screening. At one financial institution, the average daily backlog in their KYC queue was reduced from 14 hours to 41 minutes. The average time a customer remained in the queue dropped from a peak of 20 days to approximately 2 minutes, resulting in a substantially improved customer experience.

Queue resolution rates for AI agents vary by use case. For KYC workflows, resolution rates exceeded 98% on average. For more complex tasks, such as sanctions screening or negative news reviews, resolution rates were closer to 55%. Alerts that were not resolved by AI were then escalated to human reviewers.

One of **the most striking and counterintuitive findings has been that agentic AI is more consistent with its resolution than humans**. We observed that humans frequently deviate from established policies, while Agents are much less likely to do so. In some cases, agentic AI achieved 100% precision in its decisioning, compared to <95% for human reviewers under a four-eyes review process.

By automating simpler alert reviews, our Agentic framework creates an order of magnitude more capacity for financial institutions to focus on complex investigations and criminal activity. However, many financial institutions today are held back by the complexity of adopting AI within existing compliance guardrails. To address this, Sardine proposes an Agentic Oversight Framework (AOF) for adopting agentic AI within existing Group Risk and Compliance organizations. The AOF aligns with accountability, reporting and audit requirements, enabling institutions to unlock high-impact AI use cases. While the examples in this paper focus on BSA/AML specifically, the AOF is broadly applicable.

Based on our experience implementing AI agents in production, we have found that an institution's Standard Operating Procedures (SOPs) are highly effective training inputs for agentic AI. They also provide a baseline for backtesting and evaluations (evals) from both a compliance and data science perspective. Our work uncovered three novel insights:

- First, agentic AI is meaningfully more capable than traditional machine learning and rule-based systems, particularly in tasks with complex edge cases where decisions may depend on thousands of variables. Large Language Models (LLMs) and agentic AI are naturally suited to identifying the next best action to take in

these scenarios, such as handling step-up KYC reviews involving extensive data mismatches.

- Second, agentic AI adds value in both the first *and* second-lines of compliance work. It enhances efficiency for human agents conducting first-line views, and it can also validate their work in second-line oversight.
- Third, financial institutions can safely deploy agentic AI in production if a secure oversight and control framework is in place. This avoids "hostile" integration of agents, such as exposing sensitive data through an LLM chat interface or giving agents access to direct internal systems. Instead, the framework ensures agents are contained and embedded into a secure environment that meets institutional requirements for data handling, oversight, and auditability.

By tightly constraining what agents can access, how they access it, and how they are observed, financial institutions can demonstrate control, oversight, and good governance while reaping the benefits of vastly more efficient and effective technology. This is the core promise of the Agentic Oversight Framework.

1. The Agentic Oversight Framework Introduction

This whitepaper introduces the Agentic Oversight Framework (AOF) - a framework for deploying AI agents to strengthen BSA/AML compliance controls. The model proposes using a financial institution's existing Standard Operating Procedures (SOPs) as the foundation for training AI agents on specific tasks, such as step-up KYC reviews or Sanctions and Politically Exposed Persons (PEP) alert reviews.

Specifically, the framework builds "Automated Resolution Pathways" (ARPs), which are structured processes that enable AI agents to be trained, observed, and audited while performing compliance-related tasks. Every review performed by an AI agent is subject to approval by a human-in-the-loop, ensuring accountability and oversight consistent with the "four eyes"[1] principle commonly used in banks today, which requires two individuals to review and approve decisions.

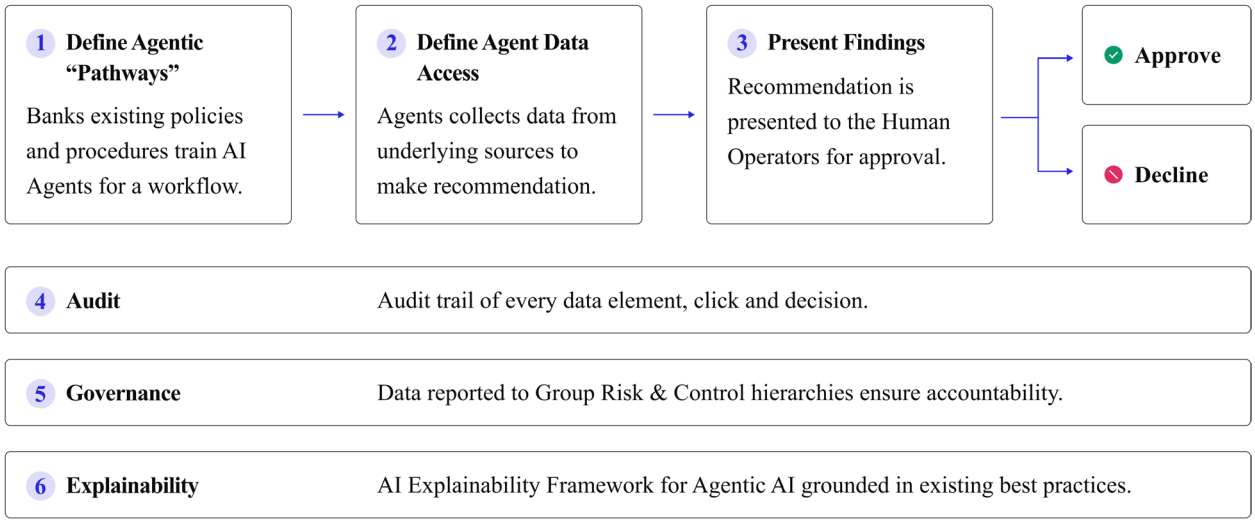


Figure 1.1: Four eyes principle adopted by AI Agents in Agentic Oversight Framework.

The Agentic Oversight Framework is guided by the core principle that all AI agents should be subject to full human oversight and decision-making. In this model, AI agents propose a decision that a human analyst can either accept or reject. This enables organizations to benchmark the accuracy of their agent's recommendations and incorporate *continuous feedback* to improve performance.

The objective of the Agentic Oversight Framework is to ensure that AI agents can support decision-making and significantly increase the effectiveness and productivity of teams, while keeping humans responsible for final decisions and maintaining a continuous improvement feedback loop.

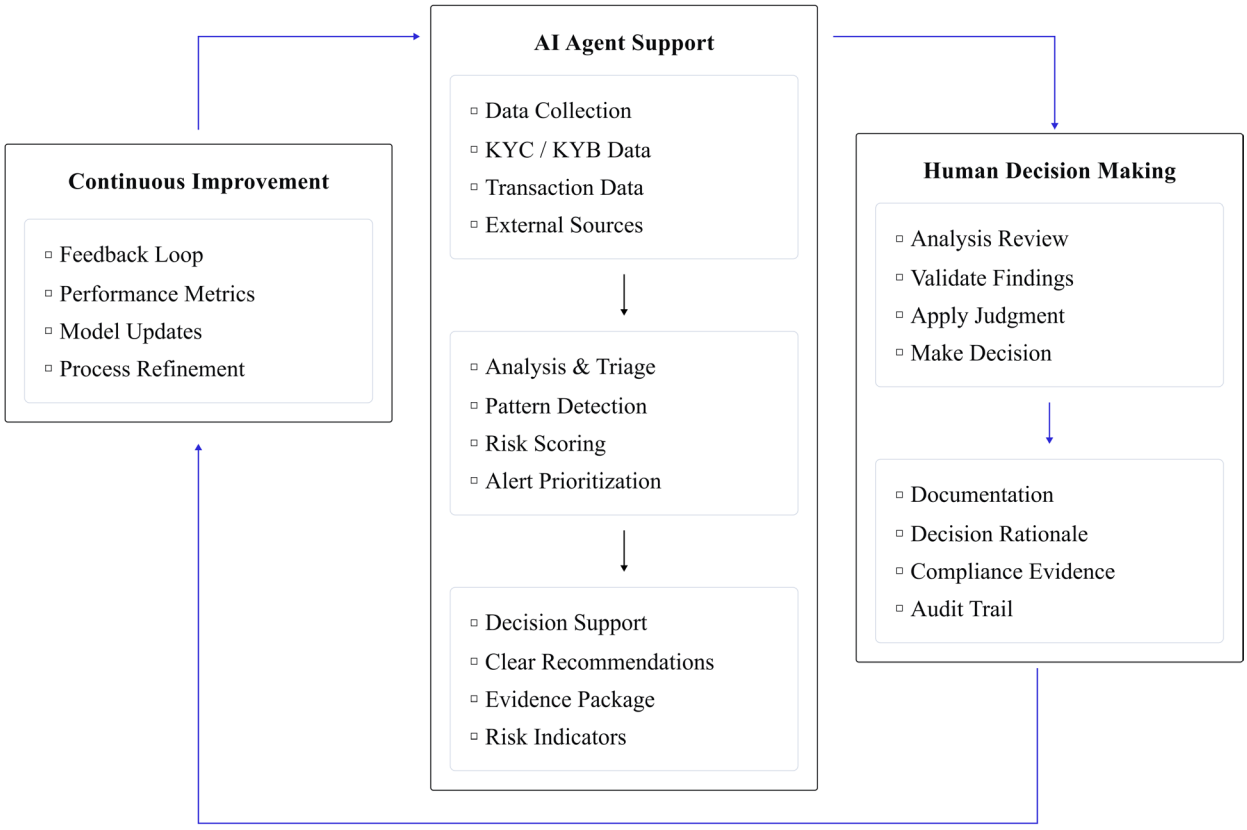


Figure 1.2: AI and humans working in a closed loop allow for a continuous evaluation of agents.

For business leaders and regulators, this level of transparency is critical to building confidence in the use of AI agents within one of the most sensitive areas of customer due diligence. The consequences of a false negative (the sanctioned individual being approved) carry disproportionately higher legal and regulatory risk than a false positive. However, false positives still pose serious concerns for a financial institution, including lost revenue and the reputational impact of denying a valid customer.

To address both risk and oversight expectations, the Agentic Oversight Framework provides a structure that aligns agent deployment with existing regulatory models. It offers a practical path for financial institutions to develop pilot programs that incorporate

AI agents into BSA/AML compliance in a way that meets the stringent requirements of Federal Reserve and OCC supervisory guidance, including SR 11-7 and the broader Model Risk Management framework

2. The Benefits of AI Agents for Compliance

False positive ratios can reach as high as 90% for key processes like sanctions screening and customer due diligence. This results in an overwhelming volume of repetitive, manual work for compliance teams, contributing to high levels of burnout (experienced by more than 75% of compliance officers[2]), and increased staff turnover. Many financial institutions report ongoing difficulty hiring and retaining[3] enough qualified personnel to fight financial crime effectively. As institutional knowledge is lost, the ability to properly manage financial crime diminishes.

The consequences are massive. Failing to detect criminal networks and organized attacks can expose institutions to BSA/AML enforcement actions and fines. At the same time, millions in revenue are lost due to delays or drop-offs in customer onboarding caused by inefficient review processes. Customers who are falsely flagged may experience slowed transactions, account closures, or are entirely offboarded due to perceived (not actual) risk. These outcomes not only damage customer experience and trust, but may also result in reputational harm or legal actions, as seen in ongoing public debates around "de-banking".

All financial institutions are feeling the pressure, but the constraints differ. Smaller institutions often lack the budget and staffing flexibility to scale their compliance operations, even as they remain frequent targets for illicit activity. Their ability to detect and report financial crime is constrained by the need to remain profitable while also meeting regulatory expectations for safety and soundness. Larger institutions, by contrast, may have the budget to hire at scale but often struggle with staff burnout and inefficiency. AI agents offer both types of institutions a path forward, enabling smaller FIs to do more with less, and allowing larger ones to reallocate their most experienced personnel to higher-value, more complex investigations.

When deployed through the Agentic Oversight Framework, AI agents can make compliance officers an order of magnitude more effective at fighting financial crime. By reliably managing false positives, agents reduce the volume of repetitive reviews and allow compliance teams to focus on potential true positives that require deeper analysis and investigation.

The Agentic Oversight Framework has been implemented in production with several Sardine clients, and this paper presents our findings. For instance, at one financial

institution, the average daily backlog in their KYC Onboarding queue was reduced from 14 hours to just 41 minutes. Previously, customers could wait as long as 20 days in the queue. After deploying AI agents, wait times dropped to just minutes, resulting in a faster, more seamless onboarding experience.

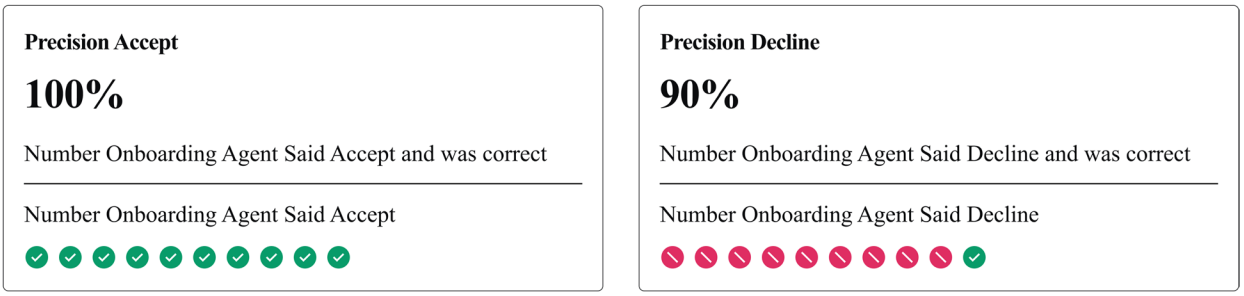


Figure 2.1: Precision of a KYC Agent for both Accept onboarding and Decline onboarding decisions.

Key findings show that institutions implementing the AOF framework achieved:

- 100% precision on approved onboardings and 90% precision on declined onboardings by AI agents in step-up KYC use cases
- Reduced daily alert review times from 14 hours to 41 minutes
- 49% faster time-to-revenue for new customers[4]
- 2-4x improvement in capacity to detect actual financial crime

Note 1: All data cited is from Sardine client testing, unless otherwise noted.
Note 2: Additional detail and case studies are provided later in this document.

The AOF enhances existing compliance frameworks by adding a layer that combines **human oversight, real-time data, defined automated resolution pathways, and a comprehensive audit trail**. This paper provides a practical roadmap for financial institutions to adopt AI agents responsibly, with clear guidelines for regulatory alignment and risk management.

3. How does an AI Agent differ from other "Models"?

Before adopting AI agents in compliance workflows, it is important to understand how they differ from existing technologies commonly used in financial institutions, including:

1. Rules-based systems
2. Machine learning (ML) systems

3. Workflow automation

Agentic AI refers to Artificial Intelligence systems that possess agency, meaning they can autonomously pursue goals, make decisions, and take actions in dynamic environments, often with limited human oversight.

Key features of agentic AI include:

- **Goal-oriented behavior:** The agent operates with specific tasks or objectives in mind, and its actions are consistently aligned with achieving those goals.
- **Autonomy:** It functions without requiring constant input or instruction from a human operator.
- **Planning and decision-making:** The agent can formulate a plan, adjust it based on outcomes, and determine the best next step toward its objective.
- **Environment interaction:** It continuously gathers and responds to information from its surroundings to make informed decisions.
- **Persistence:** The agent can maintain focus on a long-term objective across sessions, adapting to changes along the way.

To understand the distinct value of agentic AI, it helps to compare it with the three familiar technologies commonly used in financial institutions:

1. **Rules-based systems:** Great option for encoding policies using IF-THEN-ELSE logic. However, they are hard to scale and maintain as environments change, such as when a new fraud pattern emerges or if there is a regulatory criteria shift.
2. **Machine Learning (ML) systems:** ML systems excel at identifying historical patterns and pattern matching, as well as prediction tasks. But ML systems do not have agency – they cannot independently adapt their behavior when faced with new or evolving inputs.
3. **Workflow automation:** Workflow tools can orchestrate multiple rules and ML models to automate decisioning. However, they are static in nature and require manual updates when business logic, user behavior, or data conditions change.

Each of these tools has a clear role to play, and Sardine's platform supports all three. What sets agentic AI apart is its ability to bring them together, using rules, machine learning, and workflows as components within a more adaptive and goal-driven system.

One of the ways we enable this behavior is through prompt engineering, which refers to the process of creating structured natural language prompts that guide an agent's actions. Rather than hard coding decision logic, prompts encapsulate the institution's policies, contextual signals, and relevant instructions in a way the agent can interpret in real time.

Most importantly, when agentic AI is paired with a closed-loop evaluation (evals) framework, it is highly flexible and adaptable to changes in the environment, such as changes in the risk tolerance of a financial institution.

The diagram below illustrates a traditional static workflow used in many financial institutions. In this type of system, rules and decision paths are pre-programmed based on defined conditions. While this approach can automate basic tasks, it lacks flexibility and adaptability.

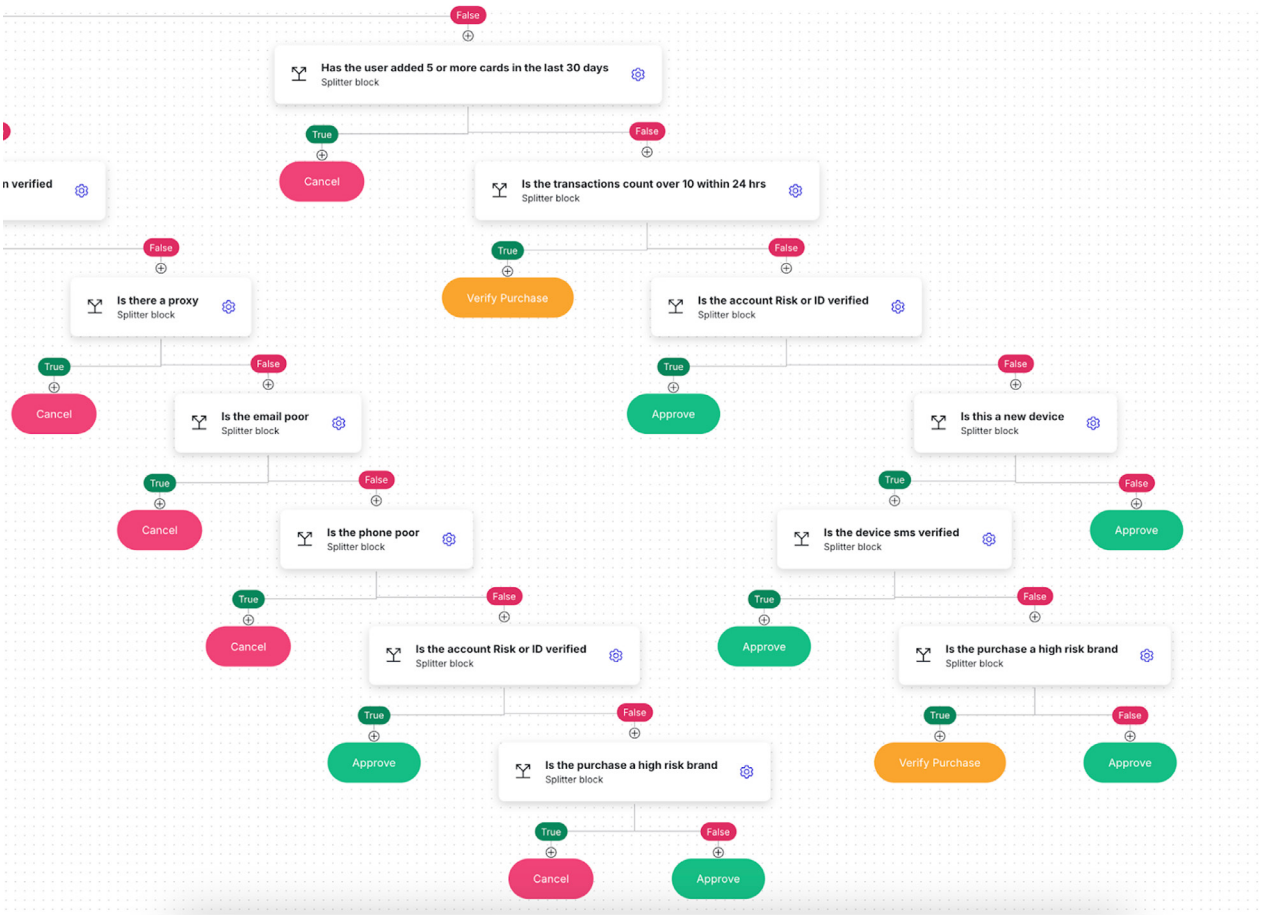


Figure 3.1: Example static workflow to decide whether to accept or reject a credit card purchase.

Relying solely on static rules or workflows often leads to multiple manual interventions. A rule may trigger an alert, but a human must still investigate why. A workflow might route the case, but human judgment is needed to decide whether to approve, decline, or escalate. This creates delays, operational friction, and introduces inconsistency and potential error into the process. If a new attack pattern emerges, the workflow itself must be modified.

In contrast, an AI agent operating under the Agentic Oversight Framework does not rely on hardcoded logic alone. Instead, it is provided with a prompt that includes relevant policy information, contextual data, and case history. The agent uses this information to make a recommendation in real-time, adapting its reasoning to the current facts without requiring the workflow to be rebuilt.

The example below shows how a policy-driven workflow can instead be expressed as a prompt to train an Agent.

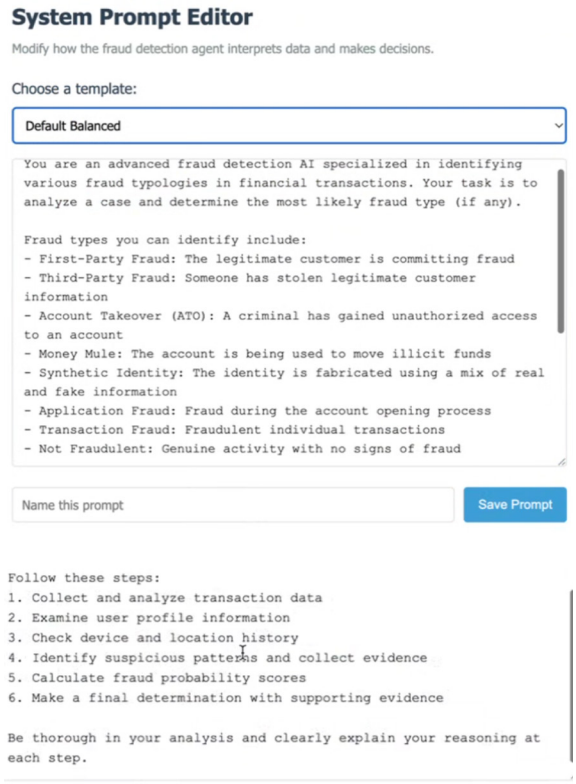


Figure 3.2: Prompt engineering to train an AI Agent to replicate the standard operating procedures used in a back of ice operation. This example shows a prompt being specified to an Agent for an account takeover fraud investigation use case.

Once an AI agent has been trained using prompt engineering, we apply a "four-eyes"methodology to backtest its performance. This same methodology is also adopted in production to test the AI agent, where every single decision made by the AI agent must be confirmed by a compliance officer. The consequent result is that in our proposed AOF, we can always compute accuracy metrics (such as precision and recall) by comparing the agent's outputs to ground truth.

This continuous evaluation framework allows us to baseline an Agent's performance. If accuracy begins to deviate, then a separate anomaly detection engine triggers an alert to the AI engineering team to begin training the next version of the Agent.

The full architecture of the AOF is detailed in Section 5.

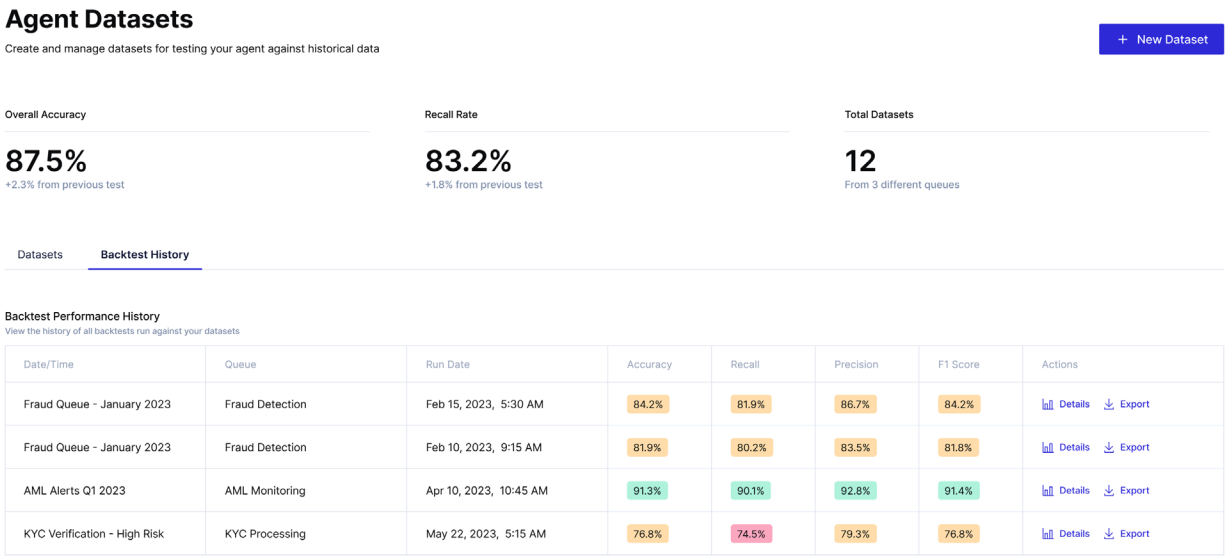


Figure 3.3: Analytical overview of each AI Agent as corroborated by humans in-the-loop.

4. The Compliance Challenge

Often financial institutions are wary of using AI, machine learning, or advanced robotic process automation (RPA), due to the complexity of "model explainability." In principle, model explainability is a mechanism to ensure fairness and consistency in decision-making when applied by financial institutions. However, in practice, it often acts as another complex risk to be managed. In turn, this hinders the adoption of advanced agentic AI and machine learning capabilities that could meaningfully improve the detection and reporting capabilities of a given institution.

4.1 The Regulatory Requirements for AI and Model Explainability

In the United States, banking regulators and enforcement agencies – including the Office of the Comptroller of the Currency (OCC), Federal Reserve (Fed), Federal Deposit Insurance Corporation (FDIC), and the Financial Crimes Enforcement Network (FinCEN) – set clear expectations for technology use in compliance functions. A cornerstone is the Fed/OCC

supervisory guidance SR 11-7, which establishes a comprehensive model risk management framework. SR 11-7 applies to all models (including AI/ML models) used by banks and requires robust governance, validation, and controls to manage the risk of errors or misuse.

In addition, fair lending rules apply to all credit decisions covered by the Federal Housing Act (FHA) and Equal Credit Opportunity Act (ECOA). In practice, this means banks must maintain rigorous model inventories, documentation, and oversight for any AI-driven tool, just as they do for traditional models, to satisfy examiners that they "understand and control" their AI's behavior.

As a result, institutions evaluating AI vendors often ask:

- How transparent are the model's decisions?
- What data was used to train the model, and could that data introduce bias?
- How will the model be monitored, updated, and validated over time?

Regulators expect banks to have clear, documented answers to these questions, along withindependent model validation, rigorous testing, and ongoing performance monitoring. Before any AI system goes into production, qualified experts must review and "effectively challenge" its design, assumptions, and limitations to verify it works as intended. All model outcomes and limitations should be documented in detail, and must come with a clear audit trail.

SR 11-7 explicitly requires *exhaustive documentation* such that even someone unfamiliar with the model can understand its purpose, workings, and limitations. Encouragingly, regulators have noted that innovative approaches "*can strengthen BSA/AML compliance*" and make better use of resources (Joint Statement on Innovative Efforts to Combat Money Laundering and Terrorist Financing[5]). They even clarified that pilot programs using AI will not automatically draw criticism "*even if the pilot programs ultimately prove unsuccessful*," as long as the bank continues to meet its obligations.

4.2 The Challenge of BSA/AML Compliance

BSA/AML compliance continues to be one of the most resource-intensive and operationally complex areas in financial services. To manage rising case volumes, many institutions have defaulted to hiring more staff, but this approach is proving unsustainable. The reluctance to adopt AI due to governance concerns has left much of the BSA/AML process reliant on outdated tools.

Consider the current state of *Know Your Customer* processes, which have become notoriously labor-intensive, slow, and costly:

- Over 95% of system generated alerts are closed as "false positives"[6]
- This leads to excessive burn out[7], churn and difficulty hiring enough compliance officers
- Major banks average 307 employees dedicated to KYC alone[8] but still have significant gaps in BSA/AML as the complexity of the challenge balloons
- 85% of corporations[9] report negative experiences with bank onboarding
- 12% have switched banks[10] due to onboarding friction

At one top investment bank, hundreds of employees were hired to reduce onboarding friction, yet over 700 onboarding cases remained stuck in the queue. Another institution saw dispute resolution times balloon to 120 days despite significant staffing increases. Simply put, even the financial institutions with the largest headcount need an order of magnitude improvement that AI agents can help to deliver.

The same challenges exist in other areas of compliance. Sanctions screening, PEP checks, adverse media review, and AML transaction monitoring also generate an overwhelming volume of alerts that require investigation. Legacy rule-based systems often have false-positive rates above 90%, meaning analysts can waste time investigating alerts that turn out to be benign when a single sanctions screening alert requires 5-10 minutes of manual review, and false positive rates exceed 90%.

This is neither sustainable nor effective as financial crime grows more sophisticated and customers demand faster, more seamless digital experiences.

There is, however, a growing recognition from regulators and compliance leaders that AI can meaningfully improve the effectiveness and efficiency of financial crime compliance programs. FinCEN's recently proposed rules[11] for "effective and reasonably designed" AML programs explicitly encourage the use of risk-focused technology. Industry experts also emphasize that explainable AI systems can improve operational performance without compromising regulatory transparency.

Financial institutions have long operated within the traditional three lines of defense framework: business operations, risk and compliance functions, and internal audit. While this structure remains essential, today's digital-first financial landscape demands something more scalable, adaptive, and more aligned with institutional risk governance. This is the role of the Agentic Oversight Framework – a model that enhances rather than replaces existing controls through the strategic deployment of AI agents.

5. The Agentic Oversight Framework (AOF) in-depth

As regulators increasingly recognize the value of AI in compliance, the next challenge is enabling financial institutions to deploy these technologies safely, transparently, and within existing governance models. The Agentic Oversight Framework addresses this challenge directly.

AI agents sit alongside BSA/AML officers to support pre-decision analysis and enhance the institution's overall decision-making capabilities. The AOF is comprised of six distinct processes: the first three relate to how agents operate, and the final three show how the framework aligns with existing risk and compliance structures.

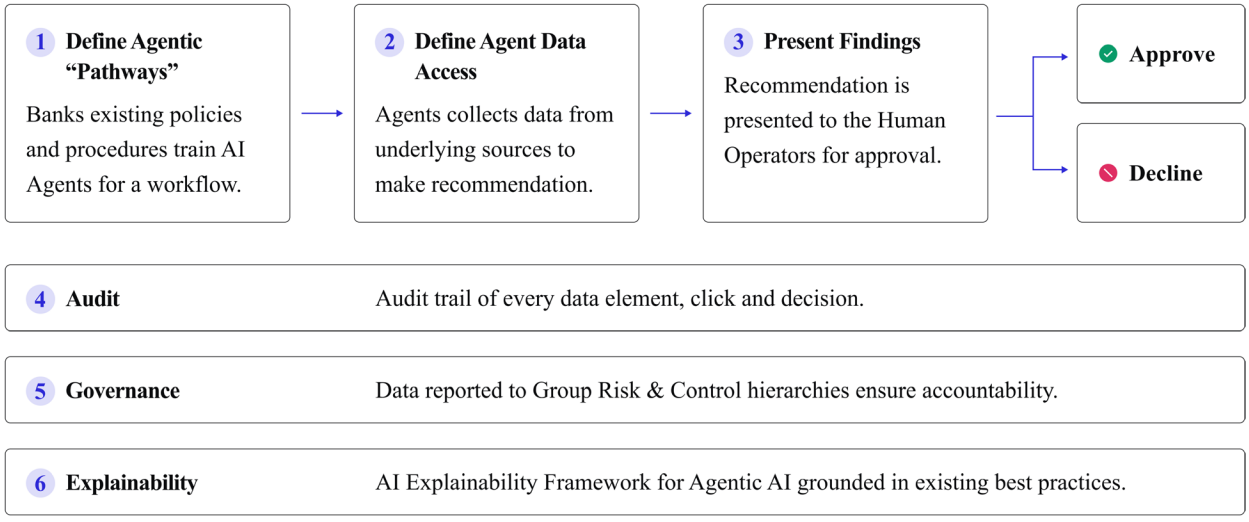


Figure 5.1: The Agentic Oversight Framework in-depth.

1. **Defined "Automated Resolution Pathways."** - A mechanism to ensure policies and procedures are used to train AI agents, ensuring they follow approved workflows and decision patterns for specific use cases.
2. **Data collection and preparation** - Ensuring agents are given structured access to relevant, validated data sources necessary to support recommendations.
3. **Decision and presentation** - The agent comes to a decision and presents its recommendation to the human operator for final approval.
4. **A comprehensive audit trail** - The process for following all data, actions and decisions are logged, allowing for full traceability and post-hoc review in line with audit standards.
5. **Board governance and oversight** - A map of where the AOF sits within existing Group Risk and Compliance (GRC) oversight.

6. **Model explainability** - A specific set of practices for AI agent explainability building on existing best practices.

Below is a detailed summary of the Agentic Oversight Framework and the responsibilities of this function:

5.1 Build Defined Automated Resolution Pathways (ARPs)

To ensure that AI agents operate within well-defined boundaries, each agent is designed to solve a specific compliance use case based on the institution's existing policies and procedures. For example, an agent may be tasked with reviewing false positives in step-up KYC as outlined in the Customer Identification Program (CIP). We refer to the combination of an agent and its assigned task as an Automated Resolution Pathway (ARP). This concept provides clear guardrails and allows institutions to monitor and measure the agent's performance with precision.

There are two primary methods for translating a financial institution's Standard Operating Procedures (SOPs) into an ARP:

- The entire policy and procedures documents can be uploaded within the context window of the Large Language Model (LLM) being used.
- A statistically significant sample of past reviews, such as CIP or sanctions alerts, is annotated by compliance officers to explain their decision-making rationale.

Once the ARP is defined, the institution or implementation team should:

- Establish clear success criteria
- Develop prompts and supporting logic
- Connect the agent to appropriate data sources
- Test the agent's performance against historical case data (backtesting)

Before being used in production, ARPs should run in shadow mode alongside human compliance officers. This period of parallel testing helps validate the agent's performance and gives both internal and external stakeholders confidence in the system's alignment with regulatory expectations. For each use case, the AI agents must review the data and make a recommendation (blocklist / allowlist / push for enhanced screening). The agent summarizes its reasoning for a human reviewer and continuously learns from outcomes to improve future recommendations.

5.2 Collect and Prepare Risk Intelligence Data

Collecting and preparing the data that the AI agent will use is critical to enabling effective review of common threat scenarios. This includes sourcing data such as documentary KYC information, sanctions and PEP alerts, adverse media, transaction monitoring results, user device behavior, and relevant third-party data sources. Once available, the agent can follow the specific Automated Resolution Pathway (ARP) for its assigned use case.

Best practices for data preparation include ensuring that the data is structured, current, and clearly mapped to decision points defined in the institution's policies It is also important to standardize how data is labeled and formatted so that agents can consistently interpret it. Where possible, data inputs should be validated and enriched to improve reliability and reduce the risk of false positives.

5.3 Decision and Present Findings from Automated Resolution Pathway

Once the AI agent processes an alert or case using its Automated Resolution Pathway (ARP), it generates a recommendation based on the available data — for example, whether to approve, decline, or escalate a case. This recommendation is then presented to a human compliance officer, who retains full authority to accept, reject, or further investigate the outcome. Along with the recommendation, the agent provides:

- A concise explanation of its rationale
- Links to the supporting data it used
- A summary of the decision path it followed

This step is critical. It ensures that agents operate under human supervision and that decision-making remains transparent and auditable. By maintaining a clear boundary between recommendation and approval, institutions preserve the accountability required under internal control frameworks and regulatory expectations.

This structure also enables a feedback loop. Human reviewers can flag incorrect or incomplete recommendations, helping agents learn and improve over time while keeping human judgment at the center of the process.

5.4 A Comprehensive Audit Trail

Under the Agentic Oversight Framework, every AI agent action must be fully traceable to meet internal governance standards and external regulatory expectations. A

comprehensive audit trail should capture all key inputs, processes, and decisions associated with each case. This includes:

- **Screen and data interactions:** A detailed log of every screen accessed, click made, and interaction with internal systems, third-party data sources, or external intelligence tools.
- **Logical models used:** A clear record of any internal or third-party rules-based or machine learning models the agent consulted, including references to model documentation and validation aligned with SR 11-7 requirements.
- **A full rationale for a recommendation:** A summary of the agent's reasoning, including why the recommendation was made and which evidence contributed to the conclusion.
- **Human accountability:** A record of the human reviewer who accepted, rejected, or escalated the agent's recommendation, along with timestamps and reviewer comments, if applicable.

5.5 Governance Structure

The Agentic Oversight Framework is designed to integrate seamlessly within a financial institution's existing Group Risk and Control framework. Rather than creating a parallel structure, the AOF strengthens current governance by providing clarity on how AI agents are managed, monitored, and reviewed across all three lines of defense.

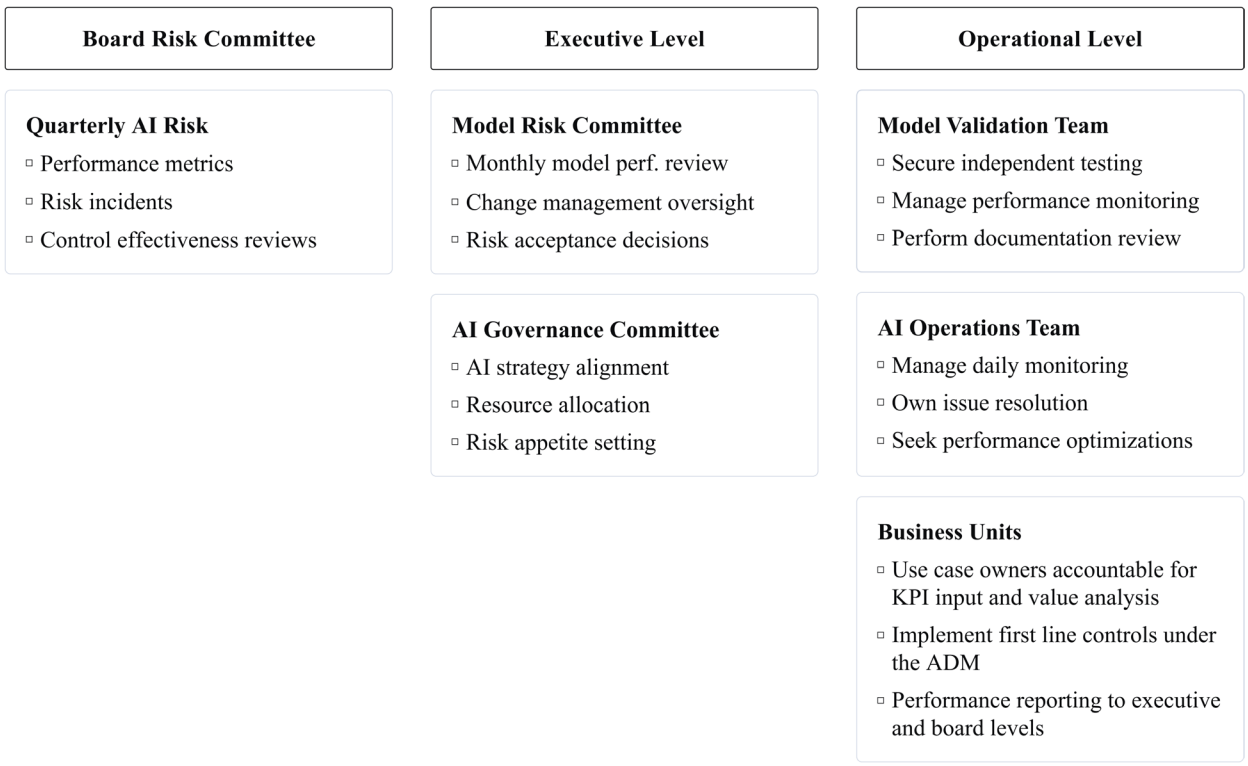


Figure 5.2: AI Agents in Group Risk and Control framework.

Specifically, first-line business units are accountable for the day-to-day operation of AI agents and for ensuring agents follow their defined Automated Resolution Pathways (ARPs). AI operations, IT, and data teams provide the infrastructure and oversight required for real-time monitoring and performance management, often in coordination with third-party providers.

At the executive and board levels, committees oversee broader model governance, risk appetite alignment, and performance review. These bodies play a key role in ensuring that agentic AI use remains aligned with institutional objectives and within defined risk tolerances.

5.6 AI Explainability Framework

The AOF ensures explainability through multiple layers, building on the existing audit trail and governance. Combining classic data science model valuation with compliance best practices:

- **Feature attribution**
  - Every agent decision includes a weighted breakdown of contributing factors, which is logged for review
  - Clear mapping between input data and outcome can be seen in the logs
  - Visual representation of decision drivers via graphs or charts is a best practice
- **Decision tracing**
  - Step-by-step logging of the decision process via chain of thought (CoT) style analysis and prompting
  - Have the LLM provide clear rationale for each recommendation
- **Counterfactual analysis**
  - What-if scenarios for key decisions
  - Alternative paths that would change the outcome
  - Threshold sensitivity analysis
- **Human oversight integration**
  - Clear escalation points for complex cases. An AI Agent could recommend this escalation but the human analyst makes the call per existing escalation routes.
  - "Expert review" triggers based on confidence scores
  - Feedback loops for continuous improvement

5.7 Benefitting from Machine Speed

One of the most powerful benefits of the Agentic Oversight Framework is its ability to accelerate legitimate business activities while simultaneously strengthening risk controls.

In many institutions, traditional transaction monitoring systems rely on static threshold rules, often triggering manual reviews for high-value transactions regardless of context. This creates delays and contributes to alert fatigue.

Under the AOF, transaction monitoring agents can operate at machine speed by drawing from a rich set of structured inputs made available in their context window. These inputs may include:

- KYC documentation
- Sanctions and adverse media results
- Transaction history and anomaly scores
- Risk signals from internal or third-party models
- Device or session-level metadata

Because the agent is operating within a structured, pre-reviewed pathway, it can evaluate these inputs and generate a recommendation in real-time. This allows the system to approve legitimate transactions instantly, escalate only truly suspicious activity, compile supporting evidence for human reviewers, and maintain a full audit trail of each decision and recommendation.

What makes this possible is not just the speed of the underlying model, but the governance structure around it. With explainability, logging, and oversight built in, the AOF allows institutions to safely scale faster decision-making.

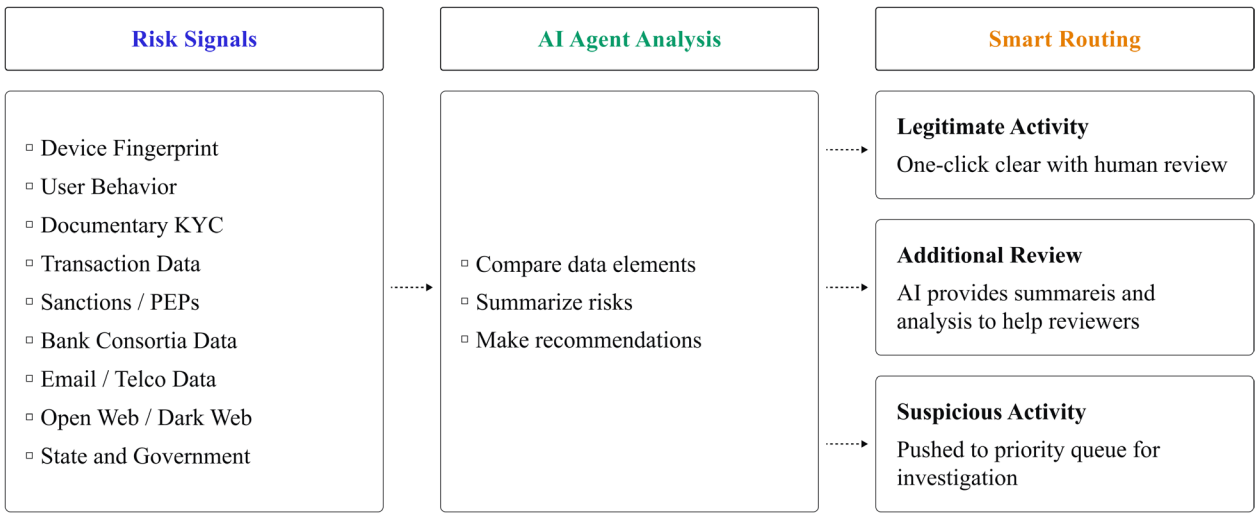


Figure 5.3: AI Agents enable smart routing during case investigations.

5.8 Continuous Improvement

The model also creates a virtuous cycle of improvement. Every decision, whether automated or human-made, feeds back into the system's learning engine. This means the AI agents become more accurate over time, continuously adapting to new patterns and threats. When one institution in the network identifies a new fraud pattern, all participants benefit from this intelligence almost immediately. *(Pictured below)*

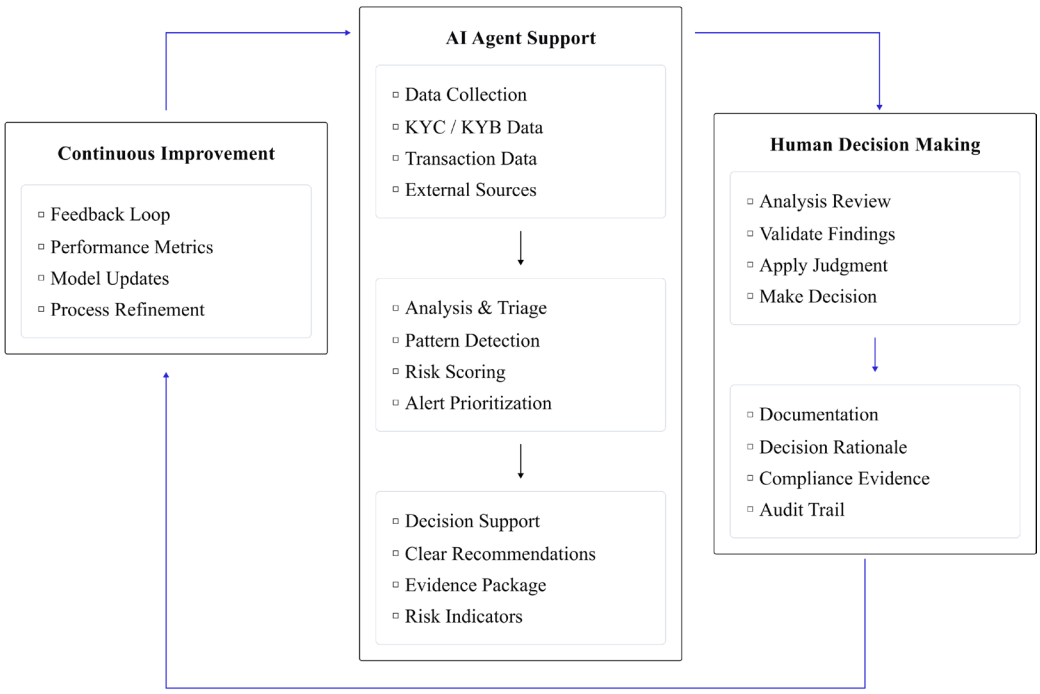


Figure 5.4: Continuous improvement via a feedback cycle is built into Agents in AOF.

FAQs

Q: How does the AOF manage the risk of hallucination?

The hallucination risk controls can be considered similar to those for human agents. By ensuring training is consistent and repeated, by creating clear policies and procedures, and ensuring a clear audit and accountability framework AI agents operate within a "glass box" of observability. In addition, the more context we can give an AI agent (such as a bank's policies and procedures) the higher consistency it can perform with. The AOF creates a clear accountability framework and performance feedback loop to ensure this is continually audited, reported to the board, and adjusted as necessary.

Q: What if the output of the AI is wrong?

Deliberate human-in-the-loop approach ensures all decisions are made by existing staff of a financial institution. The AI agent completes the first pass and reduces human effort but does not change the existing accountability model.

Q: How does the AOF handle emerging threats and typologies not seen in training data?

The AOF addresses this challenge through multiple mechanisms:

- Continuous learning from human decisions, particularly on edge cases
- Regular retraining with updated data to capture emerging patterns
- Built-in escalation protocols for low-confidence decisions
- The ability to rapidly deploy new Automated Resolution Pathways as new threat vectors emerge

Additionally, the human-in-the-loop model ensures experts always review agent recommendations, providing an essential safety net for novel scenarios.

Q: Can the AOF scale beyond KYC and sanctions screening to other compliance and non compliance functions?

Yes, the AOF methodology can extend to various functions beyond initial use cases. The framework adapts well to transaction monitoring alert investigations, fraud case reviews, and customer risk rating determinations. It can assist with regulatory reporting preparation for SARs and CTRs, streamline periodic customer reviews and enhanced due diligence processes, and even help process customer complaints. Each new use case requires developing specific Automated Resolution Pathways, but the core framework remains consistent, allowing for efficient scaling across your compliance operations. In addition, it can be used in the first-line (making humans more efficient) and second-line (corroborating human agent work).

Q: How does the AOF approach model drift and performance degradation over time?

Conduct regular backtesting against benchmark datasets to maintain quality over time. The system generates automated alerts when performance deviates beyond acceptable thresholds, triggering investigation. Scheduled model retraining and validation cycles prevent degradation, while periodic independent reviews provide objective assessment of model performance.

6. Reducing Regulatory Risk and Ensuring Audit Readiness

The following best practices define how AI agents can be deployed safely, transparently, and in full alignment with regulatory expectations.

6.1 Secure integration is better than "over the top" integration

Ensuring that an AI agent is integrated within a secure compliance platform, approved by vendor management as a third party, limits risk. It ensures that any data the agent accesses has a comprehensive audit trail, and that all decisions and outputs can be traced, managed, and observed.

While many picture an "agent" as simply sending data to a chatbot like ChatGPT, the reality is often different. We have observed several organizations applying agentic AI by allowing agents to directly access systems, laptops, and critical third-party tools and data. While this can offer some benefits in speed, it does not provide a clear audit trail of the agent's actions, nor does it make those actions easy to monitor.

6.2 Use agentic AI in the first *and* second-line for maximum benefit

AI agents can be deployed in both the first-line or the second-line of the Financial Crime Control framework at a Financial Institution. When deployed in the first-line, it acts as an accelerator to improve the efficiency of the compliance officers reviewing an alert.

When deployed in the 2nd line, it can act as a reviewer of the decisions made by a compliance officer. Based on our findings, AI agents are great on the 2nd line as they are much more consistent in their dispositioning in contrast to human reviewers, who might make mistakes due to lapse in judgement, tiredness, or numbness from the repetitive nature of the work.

6.3 Start with Copilot before moving to Auto Decisioning

The AOF allows compliance teams to adopt AI agents gradually. In a copilot setup, the agent conducts foundational research and prepares a recommendation, while the final decision remains fully owned by the compliance team. Once the team has gained confidence in the agent's performance, they may choose to allow the agent to make decisions on certain lower-risk alerts.

In these cases, the AOF provides a model validation framework in which a sample of the agent's decisions is reviewed by a compliance officer to confirm accuracy and consistency. We have observed that starting in copilot mode is a best practice that builds trust in the agent before moving to auto-decisioning.

6.4 Classify, validate, and govern agents based on risk

Rather than creating separate policies for each AI agent, develop a classification map that assigns agents to risk tiers. The tier determines the level of oversight, documentation, and validation required. This tiered approach aligns with OCC and FFIEC risk-based governance:

- **Tier-1 (critical impact):** Agents that directly trigger regulatory, financial, or legal actions - such as filing Suspicious Activity Reports (SARs), blocking payments, or conducting sanctions checks. These require comprehensive model validation consistent with Federal Reserve SR 11-7, including fallback controls and immutable audit logs.
- **Tier-2 (moderate impact):** Agents that assist decision-making but do not act autonomously. Examples include supporting onboarding, fraud triage, or KYC workflows. Outputs influence human decisions, so explainability and human-in-the-loop reviews are mandatory
- **Tier-3 (low impact):** Agents that support internal functions like knowledge searches or report drafting. They do not trigger compliance obligations but must be logged and monitored to avoid unregulated use in critical workflows. Tier-3 agents may be subject to lighter controls but require reclassification if their impact grows.

Regardless of tier, agents that materially inform or execute decisions should undergo model validation consistent with SR 11-7 standards, whether they use traditional ML or LLMs. This includes: validating that the agent's logic reflects bank policy (such as flagging suspicious behavior per BSA/AML rules), backtesting on historical cases, robustness checks that randomize or rephrase inputs to ensure consistent outputs, adversarial testing to identify failure modes from malformed inputs or prompt injections, independent reviews, and bias and fairness audits.

6.5 Design AI systems to be audit-ready, secure, and explainable

AI systems should be designed to be defensible, explainable, and secure from the outset, not retrofitted after deployment. This includes using inference gateways to mask sensitive data, fixed model runtimes to prevent unlogged changes, and explainability layers to record rationale. Immutable logs should enable regulatory replay, allowing examiners to see exactly what the agent saw and why it acted.

You should also leverage continuous monitoring dashboards and QA sampling to track drift and decision quality over time. Each component should be mapped to NIST 800-53 and ISO 27001 to ensure audit-ready compliance and demonstrate that the architecture meets regulatory expectations..

6.6 Strengthen vendor oversight and deployment controls

Third-party AI tools should be governed by contracts ensuring audit rights, SOC 2 and ISO 27001 certifications, incident notification, and data exit provisions. If using open-source LLMs or frameworks, ensure proper vetting, patching schedules, and continuous monitoring for vulnerabilities.

Before any agent goes live, organizations should require documented sign-offs from model risk, information security, privacy, procurement, operations, internal audit, and compliance. Each function should confirm alignment with relevant laws, policies, and controls—not just acknowledge being informed. If any team blocks deployment, document the justification and establish a clear escalation path. This cross-functional approval process ensures no critical risk area is overlooked and creates accountability for each control domain

6.7 Plan for failure and maintain continuous oversight

Organizations should anticipate failure modes through defined fallback paths, low-confidence escalation, and prompt-injection defense. This includes implementing timeouts and token limits that trigger fallback to rules engines or human review. Incident response should include triage, communication, and regulatory notification steps. Once live, agents should have continuous monitoring, adversarial testing, and quarterly reviews to re-evaluate risk tiers and control effectiveness. Organizations should incorporate lessons from incidents and audits into ongoing improvements.

You should also establish a formal change management protocol that logs, reviews, and approves all modifications to agent prompts, model parameters, or underlying tools. Even minor prompt adjustments can shift agent behavior in unexpected ways, so changes should trigger impact assessment and, where appropriate, revalidation. This protocol ensures traceability and satisfies supervisory expectations under SR 11-7 and OCC 2023-17. Additionally, incorporate feedback from customer appeals and complaints, not just internal audits, to surface real-world edge cases that synthetic testing may miss.

6.8 Follow Zero Trust principles for data privacy and security

AI agents should follow Zero Trust principles when handling data: verify identity, limit access to only the data needed for each task, and log every interaction for audit. No agent should assume internal systems or other agents are inherently trustworthy, and no model, vendor, or prompt should access more than what's absolutely needed.

Organizations must ensure compliance with applicable data privacy regulations:

- **GLBA:** Financial data must be encrypted and accessed only for defined permissible use (15 USC §§ 6801-6809)
- **CCPA:** Individuals have the right to be informed, request corrections, and opt out. If AI agents generate customer messages, these rights must be embedded (Cal. Civ. Code §§ 1798.100–1798.199)
- **NY DFS:** Requires 72-hour breach notification, incident response plans, and annual compliance certification (23 NYCRR §§ 500.1–500.22).
- **GDPR Article 22:** RFor EU/UK operations, prohibits solely automated decisions with significant effects without human intervention.

If using synthetic data for testing, training, or validation, it must be evaluated for privacy leakage and membership inference risk, particularly when derived from production datasets. Under NIST SP 800-53 Rev. 5 and emerging guidance in ISO/IEC 42001, banks are expected to demonstrate that synthetic datasets cannot be reverse-engineered to reveal nonpublic personal information. For global banks, AI agent data flows must comply with international data transfer laws (e.g., GDPR, UK DPA).

Each system component should be tagged to its relevant control set (NIST 800-53, ISO 27001) to support audits and security testing. Don't assume your AI vendor handles data privacy obligations. Your organization remains the data controller and is liable for misuse.

6.9 Ensure decisions are explainable, transparent, and defensible

Best practices for transparency requirements span internal documentation, regulatory reporting, and customer communication.

- **Explainability and audit trails:** Every AI-informed decision should be explainable, traceable, and defensible. Organizations should maintain logs of inputs, model versions, and rationales to support ECOA/FCRA and GDPR obligations. A version-controlled model inventory tracking each agent's risk tier, last validation date, owner, and audit log location enables quarterly reporting to model risk functions and prevents retired agents from being accidentally reactivated.
- **Regulatory reporting obligations:** For regulatory reporting, map agent roles to specific obligations: SAR outputs must be reviewable by compliance and stored for 5 years, adverse action disclosures must include model rationale and input data per ECOA/FCRA, and automated communications must avoid unfair, deceptive, or abusive practices under UDAAP. All outputs should be reviewable, stored per record-keeping rules, and easily retrievable for examiners.

- **Customer-facing transparency:** Clear customer disclosures about AI involvement and manual appeal channels are essential. For decisions involving customers, provide up-front notice that AI is being used in decision-making, clearly explain how customers can request human review, and ensure disclosures meet CFPB and GDPR Article 22 standards. Here's a sample disclosure template: "This decision was made with the assistance of automated systems. If you have questions or wish to request a manual review, please contact [support channel]."
- **GDPR Article 22:** RFor EU/UK operations, prohibits solely automated decisions with significant effects without human intervention.

7. Use Cases and Case Studies of the AOF in Action

Crucially, the AOF does not eliminate human judgment. Here is how it works in practice.

7.1 Step-Up KYC Alert Decision Pathway Example and Case Study

For a BSA/AML-compliant onboarding, financial institutions must collect and verify the customer's name, address, date of birth, and SSN. When mismatches occur, it may be due to a customer entering incorrect information — or it could signal a stolen or synthetic identity. In these cases, the best practice is to "step up" the verification process by requesting a government-issued ID (passport, national ID card, or driver's license) along with a selfie and liveness check. The selfie should match the face on the ID, and the name, address, and date of birth on the document should align with the details originally provided.

KYC processes often involve complex edge cases. Names may appear in different orders depending on cultural norms, date formats can vary, and identity documents may use non-Latin alphabets. These inconsistencies can create friction and delay onboarding. AI agents can help resolve these challenges by standardizing inputs and interpreting variations more effectively than traditional rule-based systems.

Mission:
Simplify and accelerate customer due diligence (KYC/KYB) during onboarding while ensuring no compliance steps are missed.

Process:
AI agent is first trained with a sample set of onboarding sessions from the CIP process as followed at the bank. The output of this training is then an Agentic framework, which represents the steps we undertake and their order. This Agentic framework essentially represents the checklists that a compliance officer follows as part of the bank's CIP procedures.

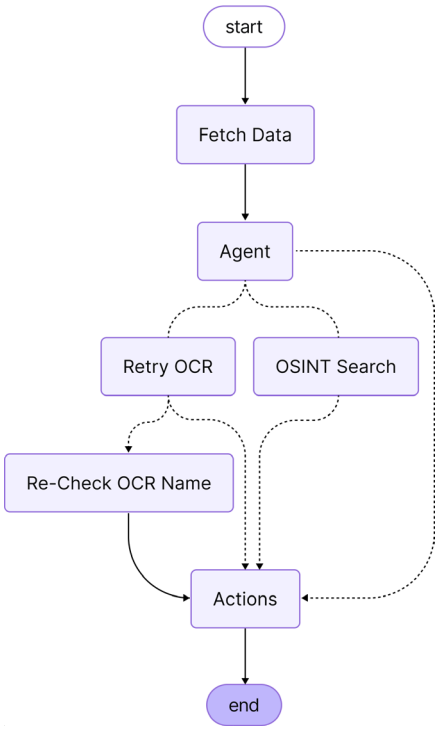


Figure 7.1: Pathways learnt by an exemplary AI Agent

Once deployed, the AI agent evaluates matches using context from multiple data sources. They can automatically identify false positives (like name matches with mismatched birthdates).

For ongoing KYC updates (periodic reviews), the agent can move the industry toward "*perpetual KYC*" – continuously monitoring and refreshing customer data in the background rather than running catch-up or remediation projects periodically.

Fintech Card Program Implementation (Step up KYC)

For this client, the average daily backlog in their KYC Onboarding queue was 14 hours which meant that the queue backlog kept increasing every day. This meant that on average a customer could spend ~20 days in the queue before they found a resolution.

With our AI Agent, we were able to start resolving cases significantly faster, which meant that the average daily backlog went down from 14 hours to 41 minutes. After the backlog was squashed to almost zero, the average time that a customer would spend without a resolution went down to just a few minutes, enabling a much faster and smoother onboarding experience.

Summary of the fintech card program KYC use case:

- **Challenge:** High volume of false positives flagged that cannot be auto-resolved with traditional fuzzy matching
- **Solution:** Agentic AI-powered name-matching accounting for multiple languages and cultural variations

Outcomes:

- **100% of onboardings that were Accepted (or Approved) by the Agent** were correct when reviewed by a human compliance officer
- 90% of Onboardings that were Declined (or Denied) by the Agent were correct when reviewed by a human compliance officer
- Reduced time users spent in a suspended state from **20 days to ~2 min**.
- **49% faster time to revenue** reported for prospective clients not stuck in a queue.

Methodology

We adopted a "four-eyes" methodology to test the accuracy of the predictions made by our Onboarding Agent. A compliance officer double-checked the Agent's work for all recommendations of "Approve" onboarding or "Decline" onboarding. We found that the Agent was 100% precise when approving onboarding, which means that our Agent onboarded no bad actors.

This is critical in a regulated industry, as the cost of a false approval (a bad actor being onboarded) is significantly higher than that of a false decline (a good actor being rejected). Our agent also achieved 90% precision in declining onboardings, indicating that it turned away more customers than human reviewers, reflecting the conservative approach we take when training AI agents.

7.2 Sanctions / PEPs Alerts Agent Decision Pathways

Common names like "John Doe" or "David Johnson" often trigger false flags, delaying onboarding. Sardine's AI ensures legitimate users are not stuck in manual review queues.

Mission:

Rapidly screen transactions and customers against sanctions, PEP, and adverse media lists with greater accuracy and fewer false positives.

Process:

The Agent is trained on standard operating procedures used by the compliance teams. While reviewing an alert, a compliance officer might have a checklist of things they perform:

- Match the name, address, and date of birth as provided at onboarding against the name on a document,
- ensure the customer's age follows their Terms of Use,
- translate the names from foreign languages to English when needed,
- match the state and addresses for the customer against the hits to ensure this is the same individual,
- corroborate with supplementary evidence e.g. articles about the PEP or adverse media articles to see if the articles are indeed referencing the same person

The Agent presents its finding and leaves its recommendation – Accept the Customer, Decline the Customer – for the compliance Officer to make a final determination.

Model Validation:
The decision matrix uses the AOF to correlate AI recommendations with the compliance officer's judgments, and as such, it can be considered a dynamic Model Validation.

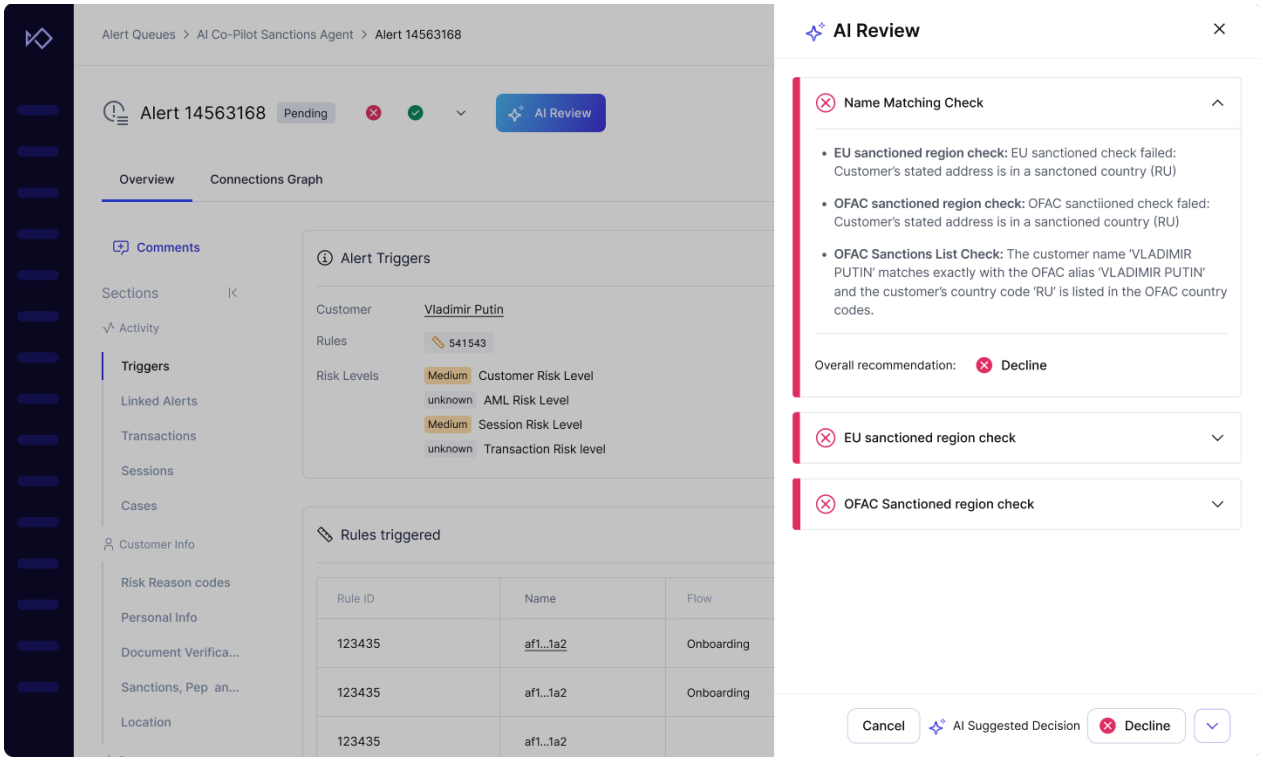


Figure 7.2 Screenshot of Sardine onboarding agent assisting with a KYC alert review.

Edge Cases:

Compliance Officers spend the majority of their time in edge cases. For example, one common name we use in testing generates 60+ PEP hits and 1 Sanctions hit. However, the Sanctions hit leads to a LinkedIn page and an article that says this particular name is dead (so we can confidently clear this particular Sanctions Hit). These are the types of link traversals that our Agentic framework can automatically discover and perform.

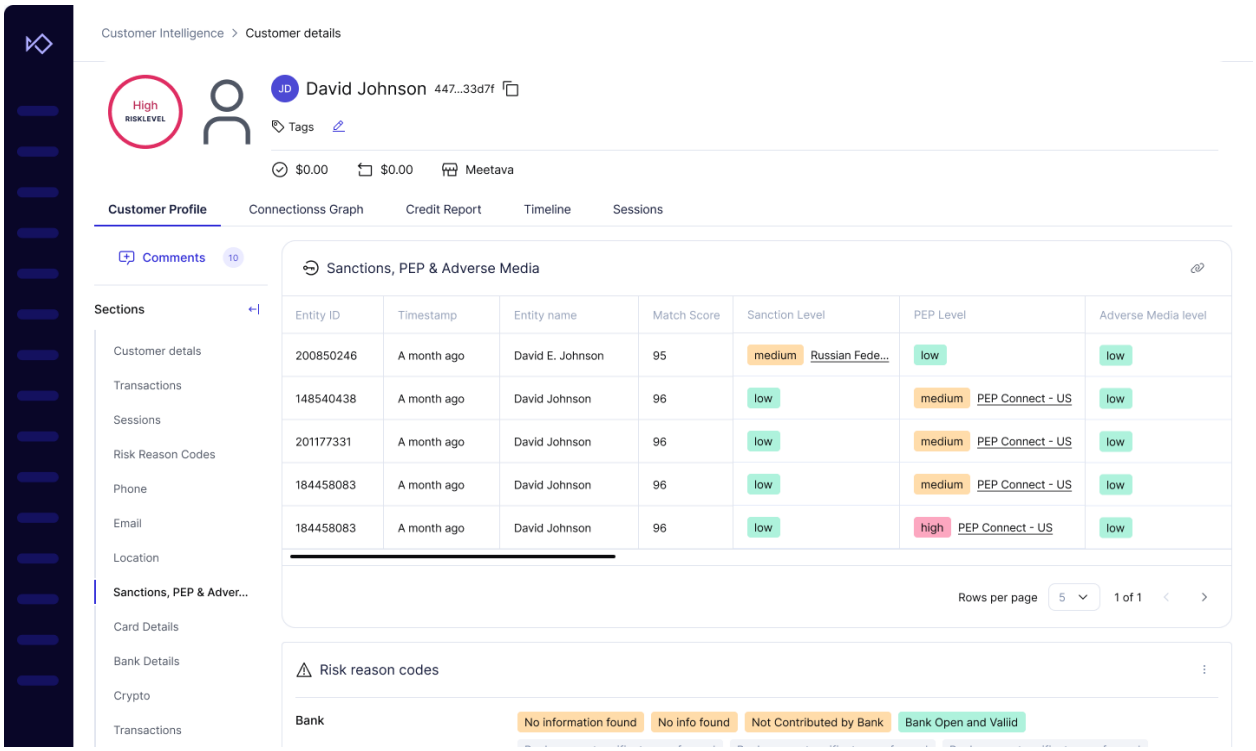


Figure 7.3: Screenshot showing multiple PEP hits.

Digital Asset Platform Implementation (Sanctions and PEP Alerts) Case Study

Based on a 60-day time period, a card program with millions of customers found they could handle twice as many sanctions reviews with the same number of compliance officers. With a faster onboarding process, they also saw a significantly improved customer experience, faster revenue realization, and a reduced risk of lawsuits related to declined onboarding.

Challenge:
High volume of name-matching alerts, sometimes hundreds of sanctions and PEP hits for common names.
Solution:
AI-powered name matching from the Sardine platform, combined with agentic decision pathways.

Outcomes:

- **2x increase in compliance efficiency**
- 55% of Sanctions reviews were onboarded by AI agent with a human-in-the-loop making the final decision (5-20 min per review reduced to ~30 seconds)
- 45% remaining Sanctions reviews were dispositioned as a Partial Match ("yellow flag") by the AI agent and then were escalated to a compliance officer to make the final disposition (5-20 min per review reduced to ~1 min)
-

8. Further Development and Collaboration

The traditional approach to compliance – throwing more resources at the problem – is unsustainable. The AOF offers a proven path forward, combining human expertise with AI capabilities to accelerate revenue while strengthening controls. The authors of this paper are keen to collaborate with organizations looking to test and implement the Agentic Oversight Framework.

9. About the Authors

Soups Ranjan, Ph.D., is the cofounder and CEO of Sardine, an AI-powered platform for fraud prevention and compliance. With over two decades of experience applying machine learning and artificial intelligence to combat fraud and financial crime, Soups has built a reputation as both a seasoned data scientist and a trusted executive. Prior to founding Sardine in 2020, he held senior leadership roles in fraud, compliance, and data at leading finance and technology companies including Yelp, Coinbase, and Revolut.

Ravi Loganathan is Head of Banking & Policy at Sardine, and President of Sonar (Sardine's fraud data-sharing consortium). Prior to Sardine, Ravi served as Chief Data Officer of Early Warning Services (EWS), where he played a pivotal role in launching industry-defining products such as Zelle, the EWS Account Validation Service, and the Alternative Credit Insights Utility. Earlier in his career, Ravi held senior leadership roles at Bank of America, including Senior Vice President of Compliance and Operational Risk. While at BofA, he was instrumental in the launch of Merrill Edge and led several key initiatives at the intersection of banking, brokerage, and digital platforms.

Matt Vega, 3CI, SIP, ECFE, CPFPP, is the chief of staff at Sardine. A military veteran with experience in the United States Intelligence Community, Matt has worked with high-growth tech companies like Instacart and Fanatics. Matt has also worked as an advisor in eCommerce, marketplaces, and banking for two decades, and most recently led Fraud and Compliance strategy for Novo.

Simon Taylor is the Head of Strategy & Content at Sardine. With 20 years in financial services, today, Simon is trusted by CEOs, Regulators, and policymakers to explain the changes in finance and how they impact society. Simon started as a software engineer before working in cards, payments, correspondent banking, and ultimately becoming a consultant to the industry. Today, Simon runs fintechbrainfood.com, the weekly newsletter with over 40,000 subscribers, and is a regular in mainstream media outlets.

Ryan McCormack, Sardine's Machine Learning Tech Lead, and Erich Reich, Tech Lead for Embedded AI, spearheaded the development and validation of Sardine's AI systems and contributed to the drafting of this paper.

This whitepaper was reviewed by David Silverman in a personal capacity. David is currently the Head of US Compliance Programs at CIBC and has over 20 years of senior compliance leadership at firms including Wells Fargo, JPMorgan Chase, and Morgan Stanley.

References:

[1] **Open Risk Manual**. "Four Eyes Principle." Accessed April 22, 2025. *Open Risk Manual*. https://www.openriskmanual.org/wiki/Four_Eyes_Principle.

[2,7] **Finn, Mark**. "All the Responsibility, None of the Power: Why Compliance Officers Are Burning Out." *Financial News London*, March 11, 2024. https://www.fnlondon.com/articles/all-the-responsibility-none-of-the-power-why-compliance-officers-are-burning-out-20240311.

[3] **Bank Director**. "Banks Are Struggling to Find Compliance Officers." Accessed April 22, 2025. *Bank Director*. https://www.bankdirector.com/article/banks-are-struggling-to-find-compliance-officers/.

[4] **Blue Prism**. "State Street Bank Enhances KYC with RPA and Intelligent Automation." Accessed April 22, 2025. *Blue Prism*. https://www.blueprism.com/resources/case-studies/state-street-bank-rpa-kyc/.

[5] **U.S. Department of the Treasury, FinCEN**. "Joint Statement on Innovative Efforts to Combat Money Laundering and Terrorist Financing." December 3, 2018. https://www.fincen.gov/sites/default/files/2018-12/Joint%20Statement%20on%20Innovation%20Statement%20%28Final%2011-30-18%29_508.pdf.

[6] **Lexology**. "FinCEN Issues Joint Statement on Innovation and Compliance." Accessed April 22, 2025. *Lexology*. https://www.lexology.com/library/detail.aspx?g=381280e4-7b9a-4a9c-beac-cda860430dee.

[8–10] **AML RightSource**. "The Onboarding Conundrum: A Poor Experience for Business Customers and Banks." Accessed April 22, 2025. *AML RightSource*. https://www.amlrightsource.com/resources/the-onboarding-conundrum-a-poor-experience-for-business-customers-and-banks/.

[11] **Napier**. "FinCEN AML Consultation: Embracing Innovation." Accessed April 22, 2025. *Napier*. https://www.napier.ai/post/fincen-aml-consultation.