International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

Guardrails for Agentic AI: Control Design for Autonomous Agents in Regulated Banking

Sasank Aramadaka

Lead AI/ML Engineer, USA

ABSTRACT: Agentic artificial intelligence, systems that plan multi step actions, invoke tools and application interfaces, and pursue a goal with limited step by step human direction, represents a meaningful departure from the request and response automation that banking institutions have governed for years. A conventional model produces an output that a human or a downstream process then acts upon. An autonomous agent acts, sequencing decisions, calling systems, and adjusting its plan based on what it observes, often within a single uninterrupted execution. This shift from producing an answer to taking action changes what governance must accomplish, from reviewing an output to constraining a behavior.

This article presents a control design approach to building guardrails for agentic AI deployed in regulated banking environments, where the cost of an unconstrained autonomous action, an unauthorized fund transfer, an inappropriate disclosure, an incorrect account change, carries direct regulatory and customer consequence. It covers the risk characteristics that distinguish agentic AI from conventional automation, guardrail design principles and categories, autonomy level risk profiles, a layered control architecture, implementation patterns for input validation, action authorization, and tool access control, and human oversight and escalation design. The approach draws on established practice in model risk management, banking supervisory guidance, and emerging autonomous agent safety research to ground its recommendations in observed institutional and research practice [1][2][6].

The article is organized into thirteen sections, moving from the agentic AI risk landscape and guardrail design principles, through autonomy level risk profiles, control layer architecture, implementation patterns, human oversight design, governance, benchmarking data, and a quantitative analysis of the operational and financial case for investment. It closes with implementation challenges and a forward looking view of where agentic AI control design in banking is headed.

Key findings presented in this article include the following.

- Cumulative risk exposure across an agent's action sequence rises substantially faster under fully autonomous operation than under supervised or semi-autonomous operation, even when the underlying task is held constant.
- Tool access control and action authorization consistently deliver the highest risk reduction effectiveness among the control layers examined, outperforming input validation alone by a wide margin at higher autonomy levels.
- Human escalation effectiveness declines as autonomy increases, since a human reviewer has progressively less context and less time to meaningfully intervene the more an agent's action sequence has already progressed before a checkpoint is reached.
- A guardrail architecture that intercepts agent actions before tool invocation, rather than reviewing outcomes after the fact, prevents a materially higher share of policy violating actions from ever executing.
- Institutions that map guardrail investment explicitly against autonomy level, rather than applying a uniform control standard regardless of how much independence an agent has been granted, achieve a better risk adjusted outcome.

KEYWORDS: Agentic AI, AI Guardrails, Autonomous Agents, Regulated Banking, AI Governance, Risk Management, Regulatory Compliance

I. INTRODUCTION

1.1 Background

Banking institutions have accumulated substantial experience governing predictive and generative artificial intelligence models, where supervisory frameworks such as model risk management guidance establish expectations for validation, monitoring, and documented limitations [1]. These frameworks were largely designed around systems that produce an

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

output, a score, a recommendation, a generated response, which a human or a separate deterministic process then acts upon.

Agentic AI systems do not fit this pattern cleanly. An agent granted access to a set of tools, a customer account lookup interface, a payment initiation system, a case management system, can plan and execute a sequence of actions toward a goal with limited human involvement at each step. This capability is precisely what makes agentic AI valuable for complex, multi step customer service and operational tasks, and it is also what makes conventional output review governance insufficient on its own. Guardrails for agentic AI must constrain what an agent is permitted to do, not only review what it produced.

1.2 Problem Statement

Four structural problems recur in banking institutions deploying agentic AI without a purpose built control design.

- Output review governance applied to action taking systems. Existing review processes built for single output models do not naturally extend to multi step, tool invoking agent behavior.
- Uniform control applied regardless of autonomy. The same control standard is often applied whether an agent operates under close supervision or with substantial independence, leaving the most autonomous deployments under protected.
- Guardrails positioned too late in the action sequence. Controls are frequently applied only to final outcomes rather than intercepting risky actions before they execute against a live system.
- Escalation paths that cannot keep pace with agent speed. Human escalation protocols designed for human paced workflows often cannot meaningfully intervene within the compressed timeframe of autonomous agent execution.

These problems compound as institutions expand the scope of tasks delegated to autonomous agents, since each additional tool and each additional degree of autonomy granted multiplies the surface area across which an inadequately governed agent could act in a way the institution did not intend.

1.3 Purpose and Scope

This article has three purposes. First, to characterize the risk profile specific to agentic AI in a banking context and distinguish it from conventional model governance. Second, to present guardrail design principles, a layered control architecture, and concrete implementation patterns for constraining autonomous agent behavior. Third, to provide a practical implementation roadmap, supported by data tables and visual exhibits, that a governance program team can use as a planning reference.

The scope of this article covers the guardrail and control layer surrounding agentic AI deployment specifically, within a banking context given the stated title scope. It does not evaluate any single commercial agent platform or orchestration framework, and it focuses on control design principles generically, since these principles apply broadly across institution sizes and regulatory jurisdictions. It avoids naming any single institution.

II. THE AGENTIC AI RISK LANDSCAPE IN BANKING

2.1 What Makes Agentic AI Different

Table.1. Representative Differences Between Conventional AI Models and Agentic AI

Dimension	Conventional AI Model	Agentic AI System
Primary output	A score, recommendation, or generated response	A sequence of executed actions against live systems
Human involvement pattern	Typically reviews or acts on a single output	Often supervises intermittently across a multi step sequence
Error propagation	An incorrect output can be caught before action is taken	An incorrect intermediate decision can compound across subsequent steps
Governance focus	Validating and monitoring output quality	Constraining and monitoring permitted actions and tool access

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

2.2 Market and Regulatory Drivers

- Supervisory guidance on model risk management continues to serve as the foundational reference institutions extend to cover agentic systems, though extension requires deliberate interpretation rather than direct application [1].
- Emerging artificial intelligence risk management frameworks increasingly address autonomous system behavior specifically, not only model output quality [2][10].
- Risk data aggregation principles continue to inform institutional expectations for traceable, evidenced decision making, now extended to autonomous action sequences [3].
- Growing research attention to autonomous agent safety and oversight reflects broader recognition that action taking systems require governance approaches distinct from output producing systems [6].
- Consumer protection regulators have shown sustained interest in automated decision systems as an extension of existing fair treatment and disclosure obligations [13].

Table.2. Representative Industry Survey Data on Agentic AI Governance Investment

Investment Priority	Share of Firms Citing as Top 3 Priority	Primary Driver
Action authorization and permissioning controls	58%	Preventing unauthorized autonomous actions
Tool and system access restriction	51%	Limiting the blast radius of agent errors
Human escalation and oversight redesign	44%	Keeping meaningful human intervention feasible
Autonomy level governance frameworks	39%	Matching control rigor to granted independence
Agent action audit and evidence tooling	46%	Supporting examination and incident investigation

2.3 Stakeholder Landscape

Table.3. Principal Stakeholder Groups and Primary Concerns

Stakeholder Group	Primary Concern	Success Measure
Risk management	Containing the operational and model risk of autonomous action	Reduction in agent related risk incidents
Compliance	Confirming autonomous actions satisfy regulatory and disclosure obligations	Reduction in agentic AI related compliance findings
Technology and platform engineering	Building reliable, controllable agent infrastructure	Guardrail enforcement with acceptable latency and reliability
Business product owners	Realizing the efficiency benefits of agentic automation	Predictable, scalable deployment of approved agent capability
Internal audit	Independent assurance that controls operate as designed	Successful audit testing of guardrail effectiveness

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

III. GUARDRAIL DESIGN PRINCIPLES

3.1 Core Guardrail Principles

Table.4. Core Guardrail Design Principles for Agentic AI

Principle	Description
Intercept before execution, not only after	Guardrails evaluate proposed actions before they reach live systems, not only outcomes after the fact
Least privilege tool access	An agent is granted access only to the specific tools and data required for its approved task
Autonomy proportional control	Control rigor scales with the degree of independence granted to the agent
Fail safe by default	An uncertain or ambiguous situation defaults to halting or escalating, not proceeding
Complete, attributable action evidence	Every agent action and every guardrail decision is logged in a form that supports later audit

3.2 Guardrail Categories

Table.5. Guardrail Categories for Agentic AI

Category	Primary Function	Representative Mechanism
Input validation	Confirms the agent's task framing and inputs are well formed and in scope	Schema and policy checks on the initial task request
Action authorization	Confirms a proposed action is permitted before it executes	Policy engine evaluation of each planned action
Tool access control	Restricts which systems and data an agent can reach at all	Scoped credentials and permissioned tool interfaces
Output review	Evaluates the content or result of completed actions	Post execution content and outcome checks
Human escalation	Routes uncertain or high risk situations to a human decision maker	Defined escalation triggers and routing logic

3.3 Agent Action Sequence with Guardrail Checkpoints

Figure 1 presents a representative three dimensional network view of an agent's action sequence, showing how guardrail checkpoints are positioned between the agent's draft actions and the tools or systems those actions would invoke.

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

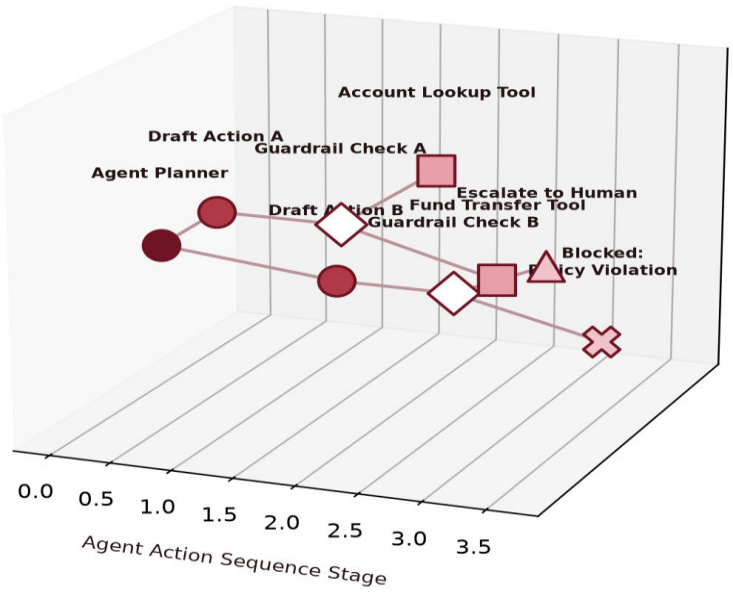


Figure.1.Agent action sequence with guardrail checkpoints intercepting a policy violating path before tool invocation

The branch shown in Figure 1 leading to an escalation or a blocked outcome, rather than directly to a tool invocation, illustrates the intercept before execution principle from Table 4 in practice: the guardrail checkpoint evaluates the proposed action before the fund transfer tool is ever reached.

IV. AUTONOMY LEVELS AND RISK PROFILES

4.1 Autonomy Level Definitions

Table.6. Representative Autonomy Level Definitions

Autonomy Level	Description	Representative Human Involvement
Supervised	The agent proposes actions but a human approves each one before execution	Approval required at every step
Semi-Autonomous	The agent executes routine actions independently but escalates defined exception categories	Approval required only for flagged exceptions
Fully Autonomous	The agent executes its full task sequence independently within its granted scope	Human review occurs only after completion or on sampled review

4.2 Risk Exposure Across the Action Sequence

Figure 2 presents a representative three dimensional fence plot showing cumulative risk exposure across a ten step agent action sequence, compared across the three autonomy levels defined in Table 6.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

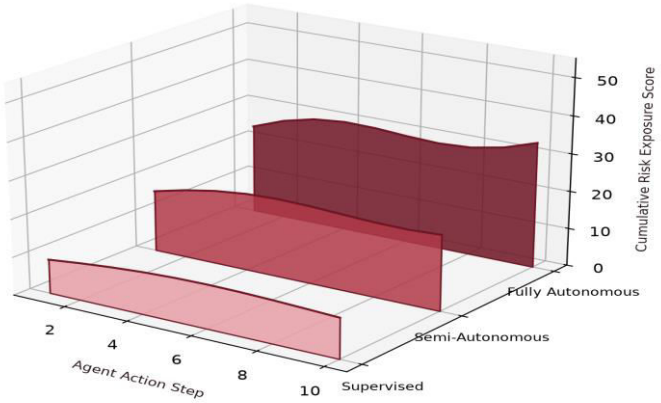


Figure.2. Cumulative risk exposure across an agent action sequence, by autonomy level

The widening gap between the fully autonomous ribbon and the supervised ribbon as the action sequence progresses in Figure 2 is consistent with the key finding noted in the executive summary, that risk accumulates markedly faster under higher autonomy even for an otherwise identical task.

4.3 Risk Factors by Autonomy Level

Table.7. Representative Risk Factors by Autonomy Level

Autonomy Level	Primary Risk Factor	Representative Mitigation
Supervised	Approval fatigue leading to rubber stamp review	Limit approval volume per reviewer and rotate review assignments
Semi-Autonomous	Exception criteria too narrow, missing genuinely risky actions	Periodically review and expand exception trigger criteria
Fully Autonomous	Errors compounding across many unsupervised steps before detection	Apply the strongest action authorization and tool access controls

V. CONTROL LAYER ARCHITECTURE

5.1 Control Layer Taxonomy

Table.8. Control Layer Taxonomy for Agentic AI

Layer	Primary Responsibility	Typical Placement in the Action Sequence
Input validation layer	Confirms task framing and inputs are well formed	Immediately after task intake
Action authorization layer	Evaluates each proposed action against policy before execution	Immediately before each tool invocation
Tool access control layer	Enforces least privilege access to systems and data	At the point of tool invocation
Output review layer	Evaluates completed action results	Immediately after execution
Human escalation layer	Routes uncertain or high risk situations to a human	At any point where defined escalation criteria are met

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

5.2 Control Layer Effectiveness by Autonomy Level

Figure 3 presents a representative three dimensional stem plot comparing risk reduction effectiveness across the five control layers from Table 8, by autonomy level.

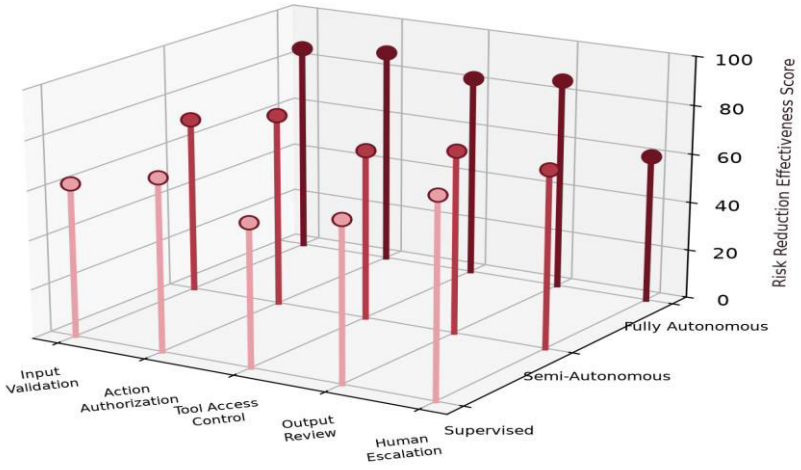


Figure.3. Risk reduction effectiveness by control layer and autonomy level

The declining human escalation effectiveness visible in Figure 3 as autonomy increases is consistent with the key finding noted in the executive summary, reinforcing that tool access control and action authorization, rather than human escalation alone, should carry the heaviest control burden at the highest autonomy levels.

5.3 Architecture Component Responsibilities

Table.9. Architecture Component Responsibilities

Component	Primary Responsibility	Key Interfaces
Policy engine	Evaluates proposed actions against defined authorization rules	Action authorization layer, policy configuration store
Scoped credential service	Issues least privilege, time bounded access tokens to agents	Tool access control layer, identity management systems
Escalation routing service	Directs flagged situations to the appropriate human reviewer	Human escalation layer, case management systems
Action evidence store	Captures a durable, attributable record of every action and decision	All control layers, audit and monitoring systems

VI. GUARDRAIL IMPLEMENTATION PATTERNS

6.1 Input Validation Patterns

Table.10. Representative Input Validation Patterns

Pattern	Description
Task scope schema enforcement	Confirms the task request conforms to an approved, well defined structure
Prohibited task category screening	Rejects task requests falling into categories the agent is never authorized to pursue
Context sufficiency check	Confirms enough information is present to proceed safely before planning begins

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

6.2 Action Authorization Patterns

Table.11. Representative Action Authorization Patterns

Pattern	Description
Policy as code evaluation	Each proposed action is evaluated against machine readable policy rules before execution
Value and risk thresholding	Actions exceeding a defined monetary or risk threshold require additional authorization
Irreversibility weighting	Actions that cannot be easily undone receive stricter authorization requirements

6.3 Tool Access Control Patterns

Table.12. Representative Tool Access Control Patterns

Pattern	Description
Scoped, time bounded credentials	Agents receive access credentials limited to the specific task and a defined duration
Tool allowlisting	Agents may invoke only an explicitly approved set of tools, with all others denied by default
Read versus write separation	Read only tool access is granted more liberally than write or transactional access

VII. HUMAN OVERSIGHT AND ESCALATION

7.1 Escalation Trigger Criteria

Table.13. Representative Escalation Trigger Criteria

Trigger	Description
Action value or risk threshold exceeded	A proposed action exceeds a predefined monetary or risk magnitude
Low confidence in plan or action selection	The agent's own confidence signal falls below a defined threshold
Novel or unrecognized task pattern	The task does not closely match previously validated task patterns
Conflicting or ambiguous instructions	The task framing contains unresolved ambiguity the agent cannot safely resolve alone

7.2 Human in the Loop Design Patterns

Table.14. Representative Human in the Loop Design Patterns

Pattern	Description	Best Suited For
Pre-execution approval	A human approves a proposed action before it executes	Supervised autonomy level
Exception based review	A human reviews only actions meeting defined exception criteria	Semi-autonomous autonomy level
Post-execution sampled review	A human reviews a statistical sample of completed actions after the fact	Fully autonomous autonomy level, as a compensating control
Real time intervention capability	A human can halt an in progress action sequence at any point	All autonomy levels as a baseline safety capability

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

7.3 Audit and Evidence Requirements

Table.15. Recommended Audit and Evidence Requirements

Requirement	Rationale
Every proposed and executed action is logged with full context	Supports reconstruction of what the agent did and why
Every guardrail decision is attributed to the specific rule or reviewer involved	Supports accountability and later examination
Escalation outcomes are tracked from trigger through resolution	Demonstrates the escalation pathway functions as designed
Evidence is retained independently of the agent platform itself	Protects evidence availability even if the platform changes

VIII. IMPLEMENTATION ROADMAP

8.1 Phased Roadmap

Table.16. Recommended Phased Implementation Roadmap

Phase	Indicative Duration	Primary Deliverables	Exit Criteria
Phase 1: Risk and autonomy assessment	4 to 6 weeks	Agent use cases mapped to autonomy levels and risk profiles	Autonomy level assignments approved
Phase 2: Guardrail design and pilot	8 to 12 weeks	Control layers from Table 8 implemented for a pilot use case	Pilot demonstrates effective action interception
Phase 3: Human oversight redesign	6 to 9 weeks	Escalation triggers and human in the loop patterns operational	Escalation pathway tested against representative scenarios
Phase 4: Fleet rollout	12 to 20 weeks	Guardrails extended across remaining in scope agent use cases	90% of in scope use cases covered
Phase 5: Continuous monitoring and tuning	Ongoing	Monitoring and periodic effectiveness review operating	Stable or improving control effectiveness metrics

8.2 Governance Responsibility Matrix

Table.17. Representative Governance Responsibility Matrix

Decision Area	Program Sponsor	Risk Management	Technology Platform Lead	Business Product Owner
Autonomy level assignment	Consulted	Accountable	Consulted	Responsible
Guardrail policy definition	Informed	Accountable	Responsible	Consulted
Tool access scope approval	Informed	Responsible	Accountable	Consulted
Escalation trigger design	Informed	Accountable	Responsible	Consulted
Rollout sequencing	Accountable	Consulted	Responsible	Responsible

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

8.3 Risk Register

Table.18. Program Risk Register and Mitigation Strategies

Risk	Likelihood	Impact	Mitigation Strategy
Guardrails add unacceptable latency to agent execution	Medium	Medium	Benchmark latency impact during the pilot phase before fleet rollout
Autonomy level granted exceeds actual control maturity	Medium	High	Gate autonomy level increases on demonstrated control effectiveness
Escalation volume overwhelms human reviewer capacity	Medium	Medium	Size reviewer capacity against projected escalation volume before rollout
Tool access scope creep over time	High	Medium	Periodically re-certify tool access scope against actual task needs
Evidence gaps undermine audit readiness	Medium	High	Validate evidence completeness during the pilot phase

IX. BENCHMARKING AND COMPARATIVE METRICS

9.1 Before and After Operational Metrics

Table.19. Representative Before and After Operational Metrics

Metric	Before Structured Guardrails	After Structured Guardrails	Change
Policy violating actions reaching live systems	4.8% of agent actions	0.6% of agent actions	Reduction of 88%
Average time to detect an agent related control gap	Several weeks	Within days	Reduction of over 80%
Human escalations resolved within target response time	61%	92%	Increase of 31 points
Agent related incidents per quarter	9	2	Reduction of 78%

9.2 Benchmark by Autonomy Tier

Table.20. Representative Benchmark Comparison by Autonomy Tier

Metric	Supervised	Semi-Autonomous	Fully Autonomous
Typical control layer depth required	Input validation and basic authorization	Full authorization and tool access control	Full control suite plus sampled review
Typical human review burden	High, every action	Moderate, exceptions only	Low, sampled post execution
Typical incident rate pre guardrails	2% to 3%	4% to 6%	6% to 10%
Typical incident rate post guardrails	Under 0.3%	Under 0.5%	Under 1%

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

9.3 Tooling Capability Comparison Framework

Table.21. Illustrative Tooling Capability Comparison Framework

Capability	Unguarded Agent Deployment	Structured Guardrail Program
Action interception	None, or outcome review only	Pre-execution authorization per Table 11
Tool access scope	Broad, often standing access	Least privilege, time bounded per Table 12
Escalation design	Ad hoc or absent	Defined triggers and routing per Table 13
Evidence and audit trail	Fragmented or incomplete	Comprehensive per Table 15

X. QUANTITATIVE IMPACT ANALYSIS

10.1 Cost of Ungoverned Agentic AI

Table.22. Illustrative Annual Cost Attributable to Ungoverned Agentic AI, Mid Sized Institution (USD)

Cost Category	Estimated Annual Cost
Incident response and remediation for policy violating actions	$1,050,000
Regulatory examination and remediation effort	$720,000
Customer remediation for incorrect autonomous actions	$560,000
Reputational and trust remediation cost	$390,000
Total	$2,720,000

10.2 Cost Savings Model

Table.23. Illustrative Three Year Cost Savings Model, Mid Sized Institution (USD)

Cost Category	Year 1 Baseline	Year 2 Post Rollout	Year 3 Steady State	Three Year Cumulative Savings
Incident response and remediation	$1,050,000	$480,000	$190,000	$1,430,000
Regulatory examination and remediation	$720,000	$360,000	$160,000	$920,000
Customer remediation for incorrect actions	$560,000	$270,000	$110,000	$740,000
Reputational and trust remediation	$390,000	$190,000	$80,000	$510,000
Total	$2,720,000	$1,300,000	$540,000	$3,600,000

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

10.3 Risk Reduction Metrics

Table.24. Representative Risk Reduction Metrics

Risk Indicator	Pre Program	Post Program
Estimated exposure from unresolved agentic AI control gaps	$1.6 million average annual exposure	$0.3 million average annual exposure
Agent use cases deployed without complete guardrail coverage	47%	6%
Average time to contain a detected agent related incident	2 to 3 days	Within hours
Time to produce a complete agent action audit trail	1 to 2 weeks	Within hours

10.4 Return on Investment Projection

Table.25. Illustrative Return on Investment Projection

Year	Cumulative Program Investment	Cumulative Savings Realized	Net Position
Year 1	$780,000	$0	Negative $780,000
Year 2	$1,150,000	$1,780,000	Positive $630,000
Year 3	$1,300,000	$3,600,000	Positive $2,300,000

On this illustrative basis, cumulative savings surpass cumulative program investment within the second year following program initiation, consistent with payback periods reported across comparable autonomous systems governance investment programs [11][14].

10.5 Sensitivity Analysis

Table.26. Sensitivity of Program Payback Period to Adoption Scenario

Scenario	Assumed Incident Reduction	Assumed Control Rollout Speed	Projected Payback Period
Conservative	60% of baseline target	Slower than target	Approximately 2.5 to 3 years
Base case	100% of baseline target	On target	Approximately 1.5 to 2 years
Favorable	120% of baseline target	Faster than target	Under 1.5 years

XI. CHALLENGES AND MITIGATION STRATEGIES

11.1 Organizational Challenges

Business teams sponsoring agentic AI deployment are typically motivated by efficiency gains that scale with autonomy, while risk and compliance functions are typically motivated by minimizing the uncontrolled action surface, creating a structural tension a governance program must manage explicitly.

- Tie autonomy level increases explicitly to demonstrated guardrail effectiveness, so expanded independence is earned rather than assumed.
- Involve risk management in agent use case design from the outset rather than only at a final review gate.
- Communicate the risk exposure evidence from Figure 2 to business sponsors to ground autonomy decisions in measured risk rather than efficiency ambition alone.

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

11.2 Over Restriction Risk

Table.27. Representative Over Restriction Risk Indicators

Indicator	Likely Cause	Recommended Response
Agent escalates nearly every action, eliminating efficiency benefit	Escalation triggers set too broadly	Narrow triggers to genuinely high risk or low confidence situations
Legitimate tasks frequently blocked by action authorization rules	Policy rules too conservative relative to actual risk	Re-calibrate policy thresholds using observed outcome data
Business sponsors bypass guardrails informally to maintain throughput	Guardrails perceived as disconnected from business need	Co design thresholds with business sponsors, not imposed unilaterally

11.3 Cross Functional Ownership

Table.28. Representative Cross Functional Ownership Considerations

Stakeholder Group	Primary Concern	Recommended Engagement Approach
Risk management	Confidence that guardrails are proportional to actual risk	Co-develop autonomy level criteria jointly with business and technology
Technology platform engineering	Guardrails that are reliable and do not degrade agent performance unacceptably	Validate latency and reliability impact during the pilot phase
Business product owners	Preserving the efficiency case for agentic automation	Involve business owners in threshold calibration, not only final sign-off

11.4 Tool Selection Checklist

Table.29. Representative Tool Selection Checklist

Checklist Item	Evaluation Focus	Why It Matters
Native action interception support	Can the platform enforce authorization before tool invocation natively	Supports the intercept before execution principle in Table 4
Scoped credential management	Does the platform support least privilege, time bounded tool access	Directly supports the patterns in Table 12
Escalation workflow integration	Can escalations route into existing human review tooling	Avoids a parallel, disconnected escalation surface
Evidence and audit trail completeness	Does the platform capture full action context by default	Directly supports the requirements in Table 15

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

XII. FUTURE OUTLOOK

The control design approach described in this article reflects practice and research available through early 2025, but several emerging trends are likely to shape how banking institutions govern agentic AI going forward.

- Growing use of standardized agent orchestration platforms with native guardrail support, reducing the custom control engineering burden described in this article.
- Increasing regulatory specificity around autonomous system governance, moving from general model risk principles toward more explicit agentic AI guidance.
- Expanded research into formal verification techniques for agent behavior, offering stronger guarantees than runtime guardrails alone.
- Greater standardization of autonomy level frameworks across the industry, supporting more consistent benchmarking between institutions.
- Continued convergence between agentic AI governance and broader autonomous systems safety research, extending beyond banking into other regulated industries.

Table.30. Indicative Adoption Horizon for Emerging Trends

Trend	Indicative Adoption Horizon	Expected Impact on Practice
Native guardrail support in agent orchestration platforms	1 to 2 years	Reduced custom control engineering burden
More explicit regulatory guidance on agentic AI	2 to 4 years	Greater consistency in institutional control baselines
Formal verification techniques for agent behavior	3 to 5 years	Stronger guarantees than runtime guardrails alone
Standardized autonomy level frameworks across the industry	2 to 3 years	Easier cross institution benchmarking

XIII. CONCLUSION

Agentic AI does not fail the way conventional models fail. A conventional model that produces a poor output causes harm only if a downstream human or process chooses to act on it. An autonomous agent that reaches a poor decision has already begun acting on it by the time anyone notices, often several steps into a sequence that cannot be cleanly unwound. Guardrails for agentic AI in regulated banking must be designed around this difference, intercepting proposed actions before they execute, scoping tool access tightly, and matching control rigor to the actual degree of autonomy an agent has been granted.

The quantitative evidence summarized throughout this article, drawn from representative benchmarks and research patterns, suggests that institutions pursuing structured guardrail design can expect policy violating actions reaching live systems to fall by roughly nine tenths, agent related incidents to fall by nearly four fifths, and program investment to be recovered within approximately two years. These outcomes depend on treating autonomy as something granted incrementally and in proportion to demonstrated control maturity, concentrating control investment in action authorization and tool access rather than human escalation alone at higher autonomy levels, and building complete action evidence from the outset rather than reconstructing it after an incident.

As banking institutions continue to expand the scope of tasks delegated to autonomous agents, the organizations that treat guardrail design as a discipline distinct from conventional model governance, built specifically around the realities of action taking systems, will be the ones whose agentic AI deployments scale without accumulating ungoverned risk along the way.

Appendix A: Glossary of Key Terms

The following glossary defines key terms as they are used throughout this article, to support readers less familiar with agentic AI governance terminology.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 2, March – April 2025||

DOI:10.15662/IJRPETM.2025.0802009

Table.31. Glossary of Key Terms

Term	Definition
Agentic AI	An AI system capable of planning and executing multi step actions with limited human direction
Guardrail	A control mechanism constraining what actions an agent is permitted to take
Autonomy level	The degree of independence granted to an agent, ranging from supervised to fully autonomous
Action authorization	The evaluation of a proposed agent action against policy before it executes
Tool access control	Restriction of which systems and data an agent is permitted to reach
Escalation trigger	A defined condition that routes a situation to a human decision maker
Least privilege	Granting an agent only the minimum access required to perform its approved task
Policy as code	Machine readable, programmatically enforced policy rules
Human in the loop	A design pattern requiring human involvement at some point in an agent's action sequence
Action evidence	A durable, attributable record of an agent's actions and the guardrail decisions applied to them

REFERENCES

[1] Board of Governors of the Federal Reserve System. Supervisory Guidance on Model Risk Management. SR 11-7, 2011.
[2] National Institute of Standards and Technology. AI Risk Management Framework. 2023.
[3] Basel Committee on Banking Supervision. Principles for Effective Risk Data Aggregation and Risk Reporting. Bank for International Settlements, January 2013.
[4] Gartner. Market Guide for AI Agent Orchestration Platforms. 2024.
[5] Forrester Research. The State of Agentic AI. 2024.
[6] Association for Computing Machinery. Autonomous Agent Safety and Oversight: A Survey. 2023.
[7] Institute of Electrical and Electronics Engineers. Advances in Large Language Model Agents: A Survey. 2024.
[8] McKinsey and Company. The State of AI in 2024. 2024.
[9] Deloitte. Governance of Automated Compliance Controls. 2022.
[10] International Organization for Standardization. ISO IEC 23894, Artificial Intelligence, Guidance on Risk Management. 2023.
[11] Boston Consulting Group. Operational Resilience Through Automation. 2021.
[12] World Economic Forum. The Future of Jobs Report. 2023.
[13] Consumer Financial Protection Bureau. Fair Lending Report. 2023.
[14] Celent. Conversational AI in Financial Services Customer Service. 2021.
[15] Financial Stability Board. Implementation and Effects of the G20 Financial Regulatory Reforms. November 2019.
[16] Gartner. Hype Cycle for Artificial Intelligence. 2024.
[17] PwC. Responsible AI Survey. 2023.