Autonomous AI Agents in Regulated Financial Services

Governance Framework for Agentic Systems
Under DORA and the EU AI Act

CRSA-1 EU Compliance Series

Ryan Fields

March 2026

Honest Framing

This specification provides a governance framework for autonomous AI agents deployed in regulated financial services environments subject to Regulation (EU) 2022/2554 (DORA) and Regulation (EU) 2024/1689 (EU AI Act). It does not guarantee that autonomous agents will not take harmful actions. Autonomous systems act without per-action human approval by design; this framework provides risk reduction infrastructure and accountability mechanisms, not behavioral safety guarantees. It operates within the Auburn honest framing principle: probabilistic risk reduction and accountability infrastructure, not the elimination of the fundamental tension between autonomy and control.

UncleBroFields@proton.me | fieldsryanchristopher@gmail.com

Contents

1 The Agentic Gap

1.1 The Deployment Reality

Financial institutions are deploying autonomous AI agents into production environments at accelerating pace. Goldman Sachs operates a multi-model AI assistant accessing OpenAI GPT, Google Gemini, Meta Llama, and Anthropic Claude within an audited environment deployed firmwide. Morgan Stanley's AI @ Morgan Stanley has achieved 98% adoption among 16,000 financial advisors, with retrieval-augmented generation over 100,000 internal research documents. JPMorgan Chase operates over 200 AI use cases in production, with pipeline architectures combining traditional models, machine learning augmentation, and generative AI document processing. Bank of America's Erica has processed 2.4 billion interactions across 45 million clients.

These deployments share a common trajectory: they began as assistive tools—drafting documents, answering questions, summarising research—and are progressively acquiring autonomy. The assistant that drafts a trade recommendation becomes the agent that executes the trade. The chatbot that answers customer questions becomes the agent that resolves customer complaints. The document processor that extracts information becomes the agent that makes the underwriting decision.

The governance frameworks under which these systems operate were designed for a different architecture. Regulation (EU) 2022/2554 (DORA) governs ICT risk for deterministic, bounded software systems. Regulation (EU) 2024/1689 (EU AI Act) governs AI systems that produce outputs—predictions, recommendations, decisions—for human consumption. Federal Reserve SR 11-7 governs models that transform inputs into quantitative outputs within defined parameters. None of these frameworks contemplates a system that autonomously pursues goals, dynamically selects tools, modifies external state, and operates across multi-step action sequences without per-action human approval.

This is the agentic gap.

1.2 Defining Agentic AI Systems

For the purposes of this specification, an **agentic AI system** ("agent") is an AI system that exhibits all five of the following characteristics:

Definition: Agentic AI System

An agentic AI system is a composed AI system that:

1. **Pursues goals autonomously.** The system receives a high-level objective and decomposes it into sub-tasks without requiring human specification of each step.
2. **Selects tools dynamically.** The system chooses which external tools, APIs, data sources, or sub-agents to invoke based on runtime assessment of the task, rather than following a fixed pipeline topology.
3. **Observes its environment.** The system receives feedback from tool outputs, external system responses, and intermediate results, and uses this feedback to adjust its subsequent actions.
4. **Plans across multiple steps.** The system maintains an action plan spanning two or more sequential steps, with the ability to revise the plan based on intermediate observations.
5. **Modifies external state.** The system can write to databases, execute transactions, send communications, update records, or otherwise alter the state of systems beyond its own internal context.

The critical distinction is between *composed AI systems* and *agentic AI systems*. A composed AI system—as defined in the CRSA-1 EU Edition—integrates multiple AI components into a pipeline, ensemble, or multi-agent architecture. An agentic AI system is a composed AI system that additionally exhibits autonomous goal pursuit, dynamic tool selection, and external state modification. All agentic systems are composed systems; not all composed systems are agentic.

A RAG pipeline that retrieves documents and generates a response is a composed system but not an agent: its topology is fixed, it does not select tools dynamically, and it does not modify external state. An AI system that receives a customer complaint, autonomously researches the account history across multiple databases, drafts a resolution, calculates a compensation amount, and initiates a payment is an agent: it pursues a goal, selects data sources dynamically, observes intermediate results, plans across steps, and modifies external state (the payment).

1.3 Three Tiers of Agent Autonomy

Not all agents present the same governance challenge. The regulatory risk scales with the degree of autonomy—specifically, with the scope of the action envelope within which the agent operates without per-action human approval. This specification defines three autonomy tiers for financial services agents.

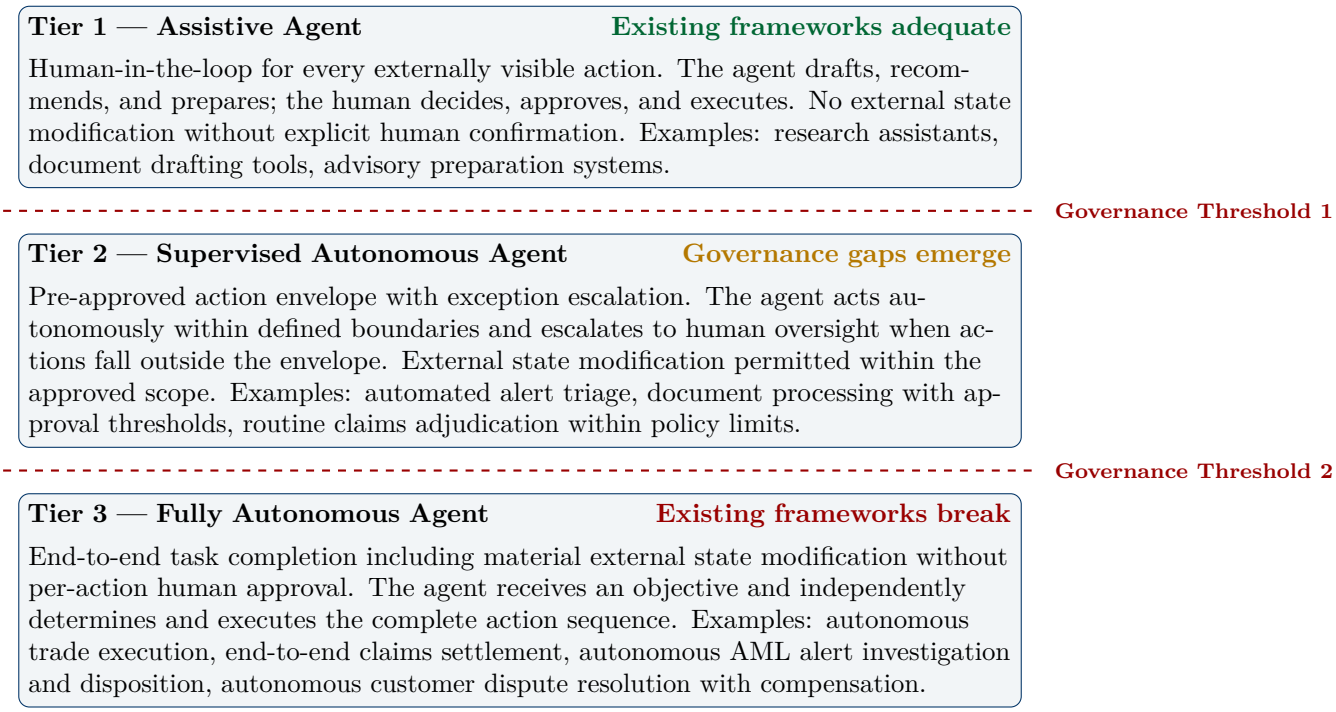

Figure 1: Three tiers of agent autonomy in financial services. Governance Threshold 1 marks the boundary where existing frameworks begin to exhibit gaps. Governance Threshold 2 marks the boundary where existing frameworks cannot provide adequate governance without the extensions defined in this specification.

Governance Threshold 1 is crossed when the agent begins to modify external state within a pre-approved envelope without per-action human confirmation. At this threshold, the institution must define the action envelope, implement monitoring for envelope violations, and establish escalation procedures. Existing model risk management frameworks (SR 11-7) and DORA's ICT risk management provisions can accommodate Tier 2 agents with extensions, but gaps appear in incident classification, resilience testing, and third-party risk management.

Governance Threshold 2 is crossed when the agent operates without a bounded action envelope—when the agent's capabilities are defined by its access to tools and data rather than by a pre-approved set of permitted actions. At this threshold, existing frameworks provide insufficient governance. The agent's risk profile is not assessable at design time because the action space is combinatorially large. Human oversight cannot operate at per-action granularity because the agent acts at machine speed. Incident classification cannot identify a discrete onset because the agent's behaviour is inherently variable. Resilience testing cannot enumerate test cases because the action space is unbounded.

1.4 The Regulatory Vacuum

No regulatory instrument applicable to EU financial institutions addresses the governance of agentic AI systems. The following table maps each relevant framework against the five defining characteristics of agentic systems.

Framework	Autonomous Goals	Dynamic Tools	Environ. Observ.	Multi-Step Plans	State Modif.
DORA (2022/2554)	—	—	—	—	—
AI Act (2024/1689)	—	—	—	—	—
SR 11-7 / OCC 2011-12	—	—	—	—	—
ECB Guide to Internal Models (2025)	—	—	—	—	—
EBA GL/2020/06 (Loan Origination)	—	—	—	—	—
BaFin AI/ICT Guidance (Dec 2025)	—	—	Partial	—	Partial
EIOPA AI Opinion (Aug 2025)	—	—	—	—	—
MiFID II Art. 17 (Algo Trading)	—	—	Partial	—	Partial
BCBS Newsletter (Mar 2022)	—	—	—	—	—
FSB AI Reports (2024–2025)	—	—	—	—	—

Table 1: Regulatory coverage of agentic AI characteristics. Red dash (—) indicates no coverage. Partial indicates tangential or implicit coverage that does not constitute a governance framework for the characteristic. No existing framework provides coverage for autonomous goal pursuit, dynamic tool selection, or multi-step planning.

BaFin's December 2025 guidance receives partial credit for environmental observation (it requires monitoring of AI outputs and anomaly detection) and state modification (it addresses data integrity and protection against unintended data changes). However, these provisions address AI systems generally, not the specific challenge of agents that observe and modify their environment

as their core operating mode. MiFID II Article 17 receives partial credit for environmental observation (algorithmic trading systems must monitor market conditions) and state modification (they execute trades), but Article 17 addresses algorithmic trading specifically, not the general class of autonomous agents operating across financial services functions.

The vacuum is total for the three characteristics that distinguish agents from other AI systems: autonomous goal pursuit, dynamic tool selection, and multi-step planning. These are not edge cases or theoretical concerns—they are the defining features of the systems that financial institutions are deploying into production today.

1.5 Scope of This Specification

This specification provides a governance framework for agentic AI systems deployed by financial institutions subject to DORA and the EU AI Act. It addresses the governance gap identified in Table 1 by providing:

1. An article-by-article analysis of where DORA's provisions fail for agentic systems and how institutions should interpret those provisions in the agentic context (Section 2).
2. An analysis of the structural paradox between the AI Act's human oversight requirements and the autonomous operation that defines agentic systems, with a proposed Agent Oversight Architecture resolving the paradox across autonomy tiers (Section 3).
3. An Agentic Risk Taxonomy (ART-1 through ART-6) defining six risk categories specific to autonomous agents in financial services, extending the Compositional Risk Taxonomy (CRT-1 through CRT-8) established in the CRSA-1 EU Edition (Section 4).
4. Four architecture-specific compliance profiles for the dominant agentic system types in financial services: autonomous trade execution, AML investigation, claims processing, and customer onboarding (Section 5).
5. Operational methodologies for the DORA Register of Information (Section 6), incident classification (Section 7), and exit strategy planning (Section 8) adapted for agentic systems.

This specification assumes familiarity with the CRSA-1 EU Edition (Fields, 2026), which establishes the compositional safety framework, obligation mapping, and risk taxonomy upon which this document builds. Where this specification extends concepts from the EU Edition, the parent concept is identified and the extension is explicitly defined.

1.6 Relationship to the CRSA-1 EU Compliance Series

This document is the third publication in the CRSA-1 EU Compliance Series:

1. **CRSA-1 EU Edition** (Published). Compositional Runtime Safety Attestation Protocol mapped to Regulation (EU) 2024/1689. Establishes the obligation mapping, compositional risk taxonomy (CRT-1 through CRT-8), five architecture profiles, and conformity assessment methodology for composed AI systems.
2. **CRSA-1 Financial Services Edition** (Forthcoming). DORA and AI Act dual-compliance framework for composed AI systems in financial services. Provides the DORA obligation mapping, financial architecture profiles, Register of Information methodology, and incident classification framework.
3. **This document.** Governance framework for the specific class of composed AI systems that exhibit autonomous agency. Extends the CRT taxonomy with agent-specific risk categories (ART-1 through ART-6), provides financial architecture profiles for agentic systems, and addresses the human oversight paradox that agents create under the AI Act.

The relationship is additive: the EU Edition provides the compositional safety foundation, the Financial Services Edition provides the DORA dual-compliance layer, and this document provides the agent-specific governance extensions. An institution deploying an autonomous credit scoring agent would reference all three documents: the EU Edition for the compositional risk framework, the Financial Services Edition for DORA compliance methodology, and this document for the agent-specific governance requirements.

CRSA-1 EU Compliance Series — Forthcoming

Subsequent editions of the CRSA-1 EU Compliance Series will provide sector-specific compliance profiles for medical devices (MDR dual-conformity) and critical infrastructure (NIS2 supply chain security). Agent-specific extensions for these sectors will be addressed as agentic deployment patterns emerge in regulated healthcare and critical infrastructure environments.

2 DORA's Implicit Agent Problem

2.1 Regulatory Basis

Regulation (EU) 2022/2554 (DORA), fully applicable since 17 January 2025, establishes the ICT risk management framework for financial entities across the European Union. DORA does not mention artificial intelligence, machine learning, algorithmic systems, or autonomous agents anywhere in its 79 articles. Its application to AI systems is entirely implicit, through technology-neutral definitions of "ICT risk" (Article 3(5): "any reasonably identifiable circumstance in relation to the use of network and information systems which, if materialised, may compromise the security of the network and information systems, of any technology-dependent tool or process, of operations and processes, or of the provision of services") and "ICT asset" (not formally defined in DORA but operationalised through Article 8(4)'s identification requirements).

BaFin's "Guidance on ICT Risks in the Use of AI at Financial Entities" (18 December 2025) is the first national supervisory document to formally confirm that AI systems are ICT assets within DORA's meaning. BaFin classifies an AI model as software —an ICT asset—and the broader AI system as "a combination of ICT assets (hardware and software) and ICT infrastructure in which a complex mathematical model is implemented." This guidance, while authoritative within Germany, does not bind other Member State supervisors, and no equivalent EU-level guidance from the ESAs addresses DORA's application to AI.

For agentic AI systems, the implicit application of DORA creates a deeper problem than for composed AI systems generally. DORA was designed for ICT systems that perform defined functions within bounded parameters. Agentic systems operate outside bounded parameters by design—their defining characteristic is autonomous action within a dynamic, potentially unbounded action space. This section analyses each relevant DORA provision and identifies where the implicit application fails, requires extension, or creates unresolvable tensions for agentic deployments.

2.2 Article 5 — Governance and Organisation

Article 5(1)–(2) — Management Body Responsibility

Text (paraphrased): The management body of the financial entity shall define, approve, oversee, and be responsible for the implementation of all arrangements related to the ICT risk management framework. The management body shall bear the ultimate responsibility for managing the financial entity's ICT risk.

Application to AI	Implicit. BaFin Guidance (Dec 2025) confirms: management body must understand AI-specific risks including model drift and hallucination.
Tier 1 (Assistive)	Manageable. The board approves the system and its use case. Human operators make all externally visible decisions. The governance model mirrors traditional software approval.
Tier 2 (Supervised)	Strained. The board approves the action envelope—the set of autonomous actions the agent may take without per-action human confirmation. The governance challenge is that the board must understand the action envelope's boundaries well enough to assess the risk it represents. For a supervised autonomous agent processing insurance claims within policy limits, the board must understand what "within policy limits" means when the agent interprets policy terms autonomously.
Tier 3 (Fully Autonomous)	**[Structural gap.]** The board approved a system whose behaviour is not fully specifiable at approval time. The agent's action space is defined by its tool access and goal specification, not by a bounded set of permitted actions. Board approval of the system does not constitute approval of each action the system may take—but Article 5 provides no intermediate governance mechanism between "approve the system" and "approve each action." The gap is not merely procedural. Article 5(2) requires the management body to "bear the ultimate responsibility for managing the financial entity's ICT risk." For a Tier 3 agent, the ICT risk is a function of the agent's runtime behaviour, which is not deterministically predictable. The management body bears responsibility for risk it cannot fully characterise at the time of approval.
Agent-Specific Requirement	Financial entities deploying Tier 2 or Tier 3 agents **SHALL** present to the management body: (a) the agent's capability envelope (the complete set of tools, data sources, and external systems the agent can access); (b) the action envelope (the subset of possible actions approved for autonomous execution); (c) the escalation boundary (the conditions under which the agent must escalate to human oversight); and (d) the risk characterisation of the residual action space between the action envelope and the capability envelope.

2.3 Article 6 — ICT Risk Management Framework

Article 6(8) — Digital Operational Resilience Strategy

Text (paraphrased): The ICT risk management framework shall include a digital operational resilience strategy setting out how the framework shall be implemented. The strategy shall include methods to address ICT risk, set ICT risk tolerance levels, and include KPIs and KRIs. Article 6(9) requires that financial entities may, as part of their digital operational resilience strategy, define a holistic ICT multi-vendor strategy showing key dependencies on ICT third-party service providers.

Application to AI

Implicit. BaFin Guidance requires documentation of pipeline topology, model dependencies, and API relationships within the resilience strategy.

Agent-Specific Problem

The resilience strategy must document the ICT reference architecture (Article 6(8)(a)). For a non-agentic composed system, the architecture is fixed at design time: Model A feeds Model B which produces an output. For an agentic system, the architecture is dynamic at runtime: the agent may invoke different tools, access different data sources, and follow different execution paths for each task invocation. The "reference architecture" of an agent is not a topology diagram—it is a capability specification describing what the agent *can* do, not what it *will* do.

Article 6(8)(c) requires KPIs and KRIs for the resilience strategy. For agentic systems, standard KPIs (uptime, response time, error rate) are insufficient. Agent-specific KRIs must capture: action envelope compliance rate (proportion of actions within the approved envelope), escalation frequency (rate at which the agent encounters situations requiring human intervention), tool utilisation distribution (whether the agent is converging on a narrow subset of tools, indicating potential concentration or behavioural drift), and autonomous decision reversal rate (proportion of agent decisions subsequently overridden or reversed).

Agent-Specific Requirement

The digital operational resilience strategy for agentic systems **SHALL** include: (a) the agent's capability specification (tools, data sources, external systems accessible) as the reference architecture; (b) agent-specific KRIs covering action envelope compliance, escalation frequency, tool utilisation, and decision reversal; and (c) the multi-vendor strategy required by Article 6(9) extended to cover tool-level dependencies, not merely model-level dependencies.

2.4 Article 8 — Identification

Article 8(4) — ICT Asset Identification and Dependency Mapping

Text (paraphrased): Financial entities shall, on a continuous basis, identify all sources of ICT risk, identify all ICT-supported business functions, roles, and responsibilities, identify the ICT assets supporting those functions, and identify all the ICT assets, including those on remote sites, network resources, and hardware equipment, and shall map those considered critical. Financial entities shall identify all dependencies on ICT third-party service providers.

Application to AI

Implicit. BaFin Guidance extends identification to "training data sets, model implementations, software libraries, hardware, and internally and externally developed software."

Agent-Specific Problem

For a non-agentic system, the ICT asset inventory is static between deployments: the components, their interconnections, and their dependencies are fixed and can be documented once and updated upon change. For an agentic system, the effective ICT asset inventory is dynamic at runtime. When an agent connects to a new MCP tool during a task execution, it has instantiated a new ICT dependency that did not exist when the inventory was last documented. The "continuous basis" requirement of Article 8(4) must be interpreted literally for agents: the inventory must reflect the agent's *current* operational state, not merely its designed state.

The dependency mapping obligation is particularly acute. An agentic system's dependencies include not only the model provider and cloud infrastructure (which are stable across invocations) but also the set of tools the agent invoked during a specific task (which vary per invocation). When an AML investigation agent queries five different external databases during one investigation and three different databases during another, the dependency profile is task-specific, not system-specific.

BaFin's guidance acknowledges "information asymmetries: not every cloud provider discloses how its AI models are developed or where data is processed." For agents, the information asymmetry is compounded: the institution may not know which external services the agent accessed during a specific task until it reviews the logs after execution.

Agent-Specific Requirement

Financial entities deploying agentic systems **SHALL** maintain: (a) a static capability inventory documenting all tools, data sources, and external systems the agent *can* access (the capability envelope); (b) a dynamic runtime inventory, updated per task invocation, documenting which tools and data sources the agent *did* access (the operational trace); and (c) an automated reconciliation mechanism verifying that the operational trace falls within the capability inventory—any access to a resource not in the capability inventory constitutes an anomaly requiring investigation.

2.5 Article 9 — Protection and Prevention

Article 9(1)–(3) — ICT Security Policies and Measures

Text (paraphrased): Financial entities shall use and maintain updated ICT systems that are reliable, have sufficient capacity, and are technologically resilient. Financial entities shall design, procure, and implement ICT security policies, procedures, protocols, and tools that aim to ensure the resilience, continuity, and availability of ICT systems, and to maintain high standards of security of data. Financial entities shall implement policies for patches, updates, and measures against intrusions and data misuse.

Application to AI

Implicit. BaFin Guidance maps protection requirements to defence against adversarial attacks, model poisoning, inference attacks, data poisoning, and prompt injection.

Agent-Specific Problem

Article 9's protection framework assumes the ICT system is the *target* of threats—external actors attempt to compromise the system. For agentic systems, the threat model inverts: the institution's own agent is an *actor* that modifies external state. The agent writes to databases, executes transactions, sends communications, and alters records. Protection must address not only threats *to* the agent (adversarial inputs, prompt injection, tool output poisoning) but also threats *from* the agent (unintended state modifications, cascading actions, scope creep beyond the action envelope).

BaFin's guidance addresses prompt injection and adversarial attacks —threats directed at the AI system. It does not address the scenario where the AI system itself is the source of unintended harm through autonomous action. A Tier 3 agent that autonomously executes a series of trades based on a misinterpreted objective is not the victim of an attack—it is performing its designed function incorrectly. The protection framework must encompass this inverted threat model.

For multi-agent systems, protection against "intrusions and data misuse" acquires an inter-agent dimension. When Agent A delegates a sub-task to Agent B, Agent B's actions are authorised by Agent A's delegation, not by human authorisation. If Agent A's delegation exceeds its own action envelope, Agent B's actions are effectively unauthorised but technically authenticated—they originate from within the system's trust boundary.

Agent-Specific Requirement

ICT security policies for agentic systems **SHALL** address: (a) the inverted threat model, including safeguards against unintended agent actions, scope creep, and cascading state modifications; (b) action envelope enforcement, ensuring the agent cannot execute actions outside its approved scope regardless of the goal specification it receives; (c) inter-agent authorisation controls for multi-agent systems, ensuring delegated actions inherit and cannot exceed the delegating agent's authorisation scope; and (d) tool output sanitisation, treating all external tool responses as untrusted input subject to validation before the agent acts upon them.

2.6 Article 10 — Detection

Article 10(1)–(2) — Anomaly Detection and Monitoring

Text (paraphrased): Financial entities shall have in place mechanisms to promptly detect anomalous activities, including ICT network performance issues and ICT-related incidents. Financial entities shall set up and implement detection mechanisms with multiple layers of control, define alert thresholds and criteria, and allocate sufficient resources and capabilities.

Application to AI

Implicit. BaFin Guidance requires continuous monitoring for model drift, adversarial attacks, anomalous outputs, and performance degradation, with "multiple layers of control" and "defined alert thresholds."

Agent-Specific Problem

Anomaly detection requires a baseline of normal behaviour against which deviations are measured. For non-agentic systems, the baseline is derivable from the system's fixed topology: normal inputs produce outputs within expected distributions. For agentic systems, behaviour is inherently variable. An agent takes different actions for different tasks by design. A customer onboarding agent that queries five databases for one applicant and seven for another is not exhibiting anomalous behaviour—it is adapting to the task.

The "alert thresholds" required by Article 10 must distinguish between three categories of behavioural variation:

Expected variability: the agent takes different actions for different inputs within its action envelope. This is normal agent operation and must not trigger alerts.

Envelope violation: the agent takes an action outside its approved action envelope. This is an anomaly requiring investigation regardless of whether the action produced a correct outcome.

Behavioural drift: the agent's action distribution shifts over time—it begins favouring certain tools, avoiding certain escalation paths, or producing systematically different outputs for similar inputs. This may not constitute an envelope violation on any individual invocation but indicates a systemic change requiring assessment.

Standard anomaly detection techniques (statistical process control, distribution monitoring) must be adapted for agents by defining baselines at the action-distribution level rather than the output-distribution level.

Agent-Specific Requirement

Detection mechanisms for agentic systems **SHALL** implement three-tier monitoring: (a) action envelope compliance monitoring, generating immediate alerts for any action outside the approved envelope; (b) behavioural distribution monitoring, tracking the agent's action patterns over time and alerting on statistically significant shifts; and (c) outcome quality monitoring, tracking the downstream effects of the agent's autonomous actions on business metrics, customer outcomes, and regulatory compliance indicators.

2.7 Article 11 — Response and Recovery

Article 11(1)–(4) — Business Continuity and Response Plans

Text (paraphrased): Financial entities shall put in place a comprehensive ICT business continuity policy, including response and recovery plans. The ICT business continuity policy shall be implemented through dedicated, appropriate, and documented arrangements, processes, and procedures.

Application to AI

Implicit. BaFin Guidance requires business continuity planning for AI systems including fallback to non-AI processes.

Agent-Specific Problem

Business continuity for a non-agentic system means restoring the system's ability to process inputs and produce outputs. Business continuity for an agentic system must additionally address the agent's mid-task state. An agent interrupted during a multi-step task may have completed some external state modifications but not others, leaving the affected systems in an inconsistent state.

This is the CRT-7 (State Corruption) risk from the CRSA-1 EU Edition, amplified for agents. In a non-agentic composed system, state corruption occurs when a component fails during processing. In an agentic system, state corruption occurs when the agent is interrupted during a multi-step action sequence that has already modified external systems. The agent may have executed a payment but not updated the ledger, approved a claim but not initiated settlement, or opened a trading position but not placed the hedge.

Article 11's response plans must address not only system restoration but also state reconciliation: identifying what the agent did before interruption, assessing the consistency of external state, and either completing or rolling back the interrupted action sequence. This requires transactional semantics across the agent's tool interactions —a capability that most current agent frameworks do not provide.

Agent-Specific Requirement

Business continuity plans for agentic systems **SHALL** include: (a) a state reconciliation procedure for identifying and resolving inconsistent external state resulting from agent interruption; (b) a rollback capability specification identifying which agent actions are reversible, which are irreversible, and the maximum rollback window for each tool interaction; (c) a graceful degradation path specifying how the agent's function is fulfilled during system unavailability, including fallback to human execution or to Tier 1 assistive mode; and (d) a mid-task handover protocol enabling a human operator to assume an interrupted agent task with full context of actions already taken.

2.8 Article 12 — Backup Policies and Procedures

Article 12(1)–(2) — Backup and Restoration

Text (paraphrased): Financial entities shall maintain and regularly test adequate backup policies, procedures, and methods for the ICT systems. Backup and restoration procedures shall not jeopardise the integrity of the network and information systems or the confidentiality of data.

Application to AI

Implicit. BaFin Guidance addresses backup of AI-related assets but acknowledges that foundation model weights accessed via API cannot be backed up by the institution.

Agent-Specific Problem

An agentic system has three categories of state that require differentiated backup treatment:

Configuration state: the agent's model, prompt templates, tool inventory, action envelope definition, and orchestration logic. This is static between deployments and can be backed up through standard procedures.

Persistent state: the agent's long-term memory, learned preferences, accumulated context, and historical action logs. For agents that maintain memory across invocations, this state grows over time and represents institutional knowledge. Loss of persistent state degrades agent performance in ways that may not be immediately detectable.

Runtime state: the agent's current task context, including the active plan, completed and pending actions, intermediate observations, and the state of external systems as modified by the agent during the current task. Runtime state is ephemeral and cannot be backed up through periodic snapshots—it must be captured continuously during execution.

Article 12's backup requirements were designed for systems where configuration state is the primary concern. For agentic systems, persistent state and runtime state present backup challenges that standard procedures do not address. Recovery time objectives must account for all three state categories, with runtime state recovery potentially requiring re-execution of interrupted tasks from checkpoint.

Agent-Specific Requirement

Backup policies for agentic systems **SHALL** address all three state categories separately: (a) configuration state backed up through standard versioned procedures with recovery point objectives aligned to deployment frequency; (b) persistent state backed up continuously or at defined intervals, with integrity verification ensuring that restored persistent state does not cause the agent to take actions based on outdated context; and (c) runtime state captured through continuous checkpointing during task execution, enabling mid-task recovery without full task re-execution.

2.9 Articles 17–19 — ICT-Related Incident Management

Article 17(1)–(3) — Incident Management Process

Text (paraphrased): Financial entities shall define, establish, and implement an ICT-related incident management process to detect, manage, and notify ICT-related incidents. Financial entities shall record all ICT-related incidents and significant cyber threats.

Application to AI

Implicit. BaFin Guidance requires incident management processes to encompass AI-specific failure modes including model drift, adversarial attacks, and data poisoning.

Agent-Specific Problem

The incident management process assumes a discrete incident with an identifiable onset. For agentic systems, three onset ambiguities arise:

Action-as-incident: when does an autonomous agent action become an "ICT-related incident"? If a Tier 3 agent autonomously executes a trade that produces a loss, is the trade an incident or a business outcome? If the agent misclassifies a customer's risk profile, is the misclassification an incident at the moment it occurs, at the moment the error is detected, or at the moment the customer suffers harm?

Cumulative onset: an agent's decisions may be individually reasonable but collectively harmful. A claims processing agent that consistently undervalues claims from a specific demographic does not produce a single identifiable incident—it produces a pattern that only becomes visible through statistical analysis over time. The "onset" is diffuse.

Delegation chain attribution: in a multi-agent system, Agent A delegates to Agent B which calls Tool C whose output causes Agent B to take a harmful action. When did the incident occur—at the delegation, at the tool call, at the action, or at the harm?

Agent-Specific Requirement

Incident management for agentic systems **SHALL** define: (a) onset criteria distinguishing between adverse agent actions (single events), cumulative behavioural patterns (statistical detection), and delegation chain failures (multi-point attribution); (b) mandatory review triggers—conditions under which the agent's autonomous actions are automatically flagged for incident assessment regardless of detected harm; and (c) attribution protocols for multi-agent delegation chains.

Article 19 — Classification and Reporting of Major Incidents

Text (paraphrased): Financial entities shall classify ICT-related incidents using the criteria in the RTS (CDR 2024/1772). Major incidents shall be reported: initial notification within 4 hours of classification (maximum 24 hours after detection), intermediate report within 72 hours, final report within 1 month.

Application to AI

Implicit. No RTS provision addresses AI-specific incident classification.

Agent-Specific Problem

The classification RTS (CDR 2024/1772) uses six materiality criteria, of which any two must be met: clients affected (>100,000 or >10%), duration (>24 hours or >2 hours for critical functions), geographical spread (≥2 Member States), data losses, reputational impact, or economic impact. Malicious unauthorised access triggers automatic major classification.

Agent-specific incidents map unevenly. A foundation model API outage maps cleanly to duration and client impact criteria. Adversarial manipulation (prompt injection causing unauthorised actions) constitutes "malicious unauthorised access," triggering automatic classification. However, two scenarios fall through the framework:

Autonomous harmful action: the agent operates within its designed parameters but produces an unintended outcome. This is not a system failure, not unauthorised access, and not a cyberattack. It is the system operating as designed but producing customer harm. None of the six criteria cleanly capture this scenario unless harm reaches impact thresholds.

Behavioural drift: the agent's decision distribution shifts gradually over weeks, producing systematically biased outcomes. There is no discrete incident to classify. The 4-hour reporting clock cannot start because there is no onset event. Classification is only possible through retrospective analysis, by which point reporting deadlines have been missed relative to the drift's onset.

Agent-Specific Requirement

Financial entities **SHALL** establish: (a) proactive classification triggers based on statistical monitoring of agent action distributions, enabling classification before individual harm events accumulate to materiality thresholds; (b) a defined attribution window specifying the maximum lookback period for determining incident onset when detected retrospectively; and (c) a classification methodology for autonomous harmful actions that do not constitute system failures.

2.10 Articles 24–26 — Digital Operational Resilience Testing

Article 24–25 — General Testing Requirements

Text (paraphrased): Financial entities shall establish, maintain, and review a comprehensive digital operational resilience testing programme including vulnerability assessments, open-source analyses, network security assessments, gap analyses, source code reviews, scenario-based tests, compatibility testing, performance testing, end-to-end testing, and penetration testing.

Application to AI

Implicit. BaFin Guidance requires resilience testing to include AI-specific scenarios including provider outage simulation and fallback testing.

Agent-Specific Problem

The testing framework assumes bounded, enumerable behaviour. An agent with access to 10 tools, each with 5 parameter configurations, operating over 5-step sequences, has $50^5 = 312.5$ million possible action paths. Exhaustive testing is impossible.

Article 25's test types create specific impossibilities for agents. *Source code review*: the agent's behaviour is not fully determined by code—the same orchestration produces different actions depending on model interpretation, tool outputs, and context. *Vulnerability assessment*: API-accessed model internals are not inspectable. *Scenario-based testing*: must include action-sequence scenarios, not merely input scenarios—testing how the agent behaves when intermediate observations deviate, tools return unexpected results, or the environment changes mid-execution. *End-to-end testing*: must cover a statistically representative sample of execution paths, not merely the happy path, exercising the agent under tool failure, ambiguous inputs, conflicting objectives, and envelope boundary conditions.

Agent-Specific Requirement

Resilience testing for agentic systems **SHALL** include: (a) action-path testing exercising representative samples of execution paths with emphasis on paths approaching the action envelope boundary; (b) adversarial scenario testing presenting inputs designed to induce actions outside the approved envelope, including indirect prompt injection through tool outputs; (c) environmental perturbation testing introducing unexpected tool failures, latency spikes, and inconsistent responses during multi-step execution; and (d) safe halt testing verifying that interruption produces a defined safe state at each stage of a multi-step sequence.

Article 26 — Threat-Led Penetration Testing

Text (paraphrased): Financial entities identified per the RTS (CDR 2025/1190) shall, at least every 3 years, carry out TLPT on live production systems. ICT third-party service providers shall participate through pooled testing arrangements per Article 26(4).

Application to AI	Implicit. The TLPT RTS identifies G-SIIs, O-SIIs, and significant payment institutions. No AI-specific methodology is prescribed.
Agent-Specific Problem	TLPT for agents must target the agent's autonomy as the attack surface. A threat actor does not merely compromise the system—they manipulate the agent into taking actions that benefit the attacker while appearing legitimate. Four agent-specific threat scenarios: *Goal manipulation*: inputs causing the agent to misinterpret its objective, executing legitimate actions toward the wrong goal. *Tool output poisoning*: compromising an external data source the agent queries, causing actions based on manipulated information. *Escalation suppression*: crafting conditions that prevent the agent from escalating to human oversight, keeping it in autonomous mode for actions requiring human approval. *Delegation exploitation*: in multi-agent systems, manipulating delegation to cause a specialist agent to take actions outside the coordinating agent's intended scope. Foundation model providers not designated as CTPPs (including OpenAI and Anthropic as of March 2026) have no Article 26(4) participation obligation. The institution must conduct TLPT on agent behaviour using the model API as-is, without provider participation.
Agent-Specific Requirement	TLPT for agentic systems **SHALL** include: (a) goal manipulation attacks; (b) tool output poisoning attacks; (c) escalation suppression attacks; and (d) delegation exploitation attacks for multi-agent systems.

2.11 Article 28 — ICT Third-Party Risk Management

Article 28(2)–(4) — Register of Information and Concentration Risk

Text (paraphrased): Financial entities shall maintain a Register of Information of all contractual arrangements with ICT third-party service providers (Article 28(3)). Financial entities shall assess whether any new arrangement may contribute to reinforcing ICT concentration risk (Article 28(4)(c)).

Application to AI

Implicit. The Register of Information ITS (CIR 2024/2956) requires registration of all ICT service provider arrangements with no de minimis threshold.

Agent-Specific Problem

For agentic systems, the third-party landscape is dynamic. An agent with access to 15 MCP tools may invoke different subsets per task. Each MCP server is a potential ICT third-party service arrangement. Three interpretive approaches exist for registration:

Tool-level: each tool connection is a separate arrangement (B.02/B.03/B.04 entries per tool). Conservative but operationally burdensome.

Provider-level: all tools from the same provider consolidated under one arrangement, B.04 covering the full tool inventory. Reduces burden but may obscure criticality differences across tools.

Function-level: tools grouped by business function, one arrangement per function. Aligns with DORA's criticality model but requires restructuring as usage evolves.

Concentration risk acquires a new dimension. When an agent autonomously selects tools, runtime preferences may create concentration the institution did not design. If the agent consistently favours one data provider's API for latency reasons, it has autonomously created a dependency the institution's concentration risk assessment must detect.

Agent-Specific Requirement

Financial entities **SHALL**: (a) register all tool providers the agent *can* access, not merely those accessed historically; (b) include tool selection logic in concentration risk assessment, monitoring for autonomous concentration; and (c) update the Register when new tools are added to the capability inventory, prior to first access. Section 6 provides the complete Tool Inventory Registration Methodology.

2.12 Articles 29–30 — Due Diligence and Contractual Provisions

Article 29(1) and Article 30(2)–(3) — Assessment and Contractual Requirements

Text (paraphrased): Financial entities shall carry out pre-contractual assessment of all relevant risks (Article 29(1)). Contracts shall include the elements in Article 30(2) for all arrangements and Article 30(3) for critical or important functions, including unrestricted audit rights (30(3)(e)(i)), TLPT participation (30(3)(d)), and exit strategies (30(3)(f)).

Application to AI	Implicit. Standard foundation model API terms do not provide DORA-compliant provisions. Cloud intermediaries offer DORA addenda.
Agent-Specific Problem	The contractual challenge extends beyond the model provider to the tool ecosystem. Each tool the agent accesses represents a potential arrangement requiring Article 30 provisions. For a Tier 3 agent accessing market data feeds, counterparty databases, payment systems, and communication platforms, the institution must negotiate DORA-compliant contracts with each tool provider. Pre-contractual assessment under Article 29 is complex for tools the agent selects dynamically. If the agent can choose between three market data providers, the institution must assess all three—even if the agent only uses one on any given day. Article 30(3)(d) requires TLPT participation for critical function providers. Agent-specific TLPT must test interaction with each tool, meaning each critical tool provider must contractually agree to participate. This negotiation burden scales linearly with the tool inventory. Exit strategies (Article 30(3)(f)) are required per critical arrangement. Tool-level exit strategies may be more achievable than model-level strategies if tools provide standardised interfaces, but the agent's response parsing and prompt engineering may be tuned to specific provider output formats, creating implicit switching costs.
Agent-Specific Requirement	Financial entities **SHALL**: (a) conduct Article 29 assessments for all tool providers in the capability inventory; (b) negotiate Article 30 provisions with each tool provider, with Article 30(3) provisions for critical function tools; (c) include agent-specific TLPT participation clauses; and (d) maintain tool-level exit strategies specifying alternative providers and migration procedures.

2.13 Summary: DORA Provisions and Agentic System Gaps

The following table consolidates the agent-specific gaps identified across DORA's provisions, mapped to the autonomy tier at which each gap becomes material. A dash (—) indicates no material gap. "Partial" indicates the gap can be mitigated through extensions to existing frameworks. **[Full]** indicates the gap cannot be resolved within DORA's existing provisions without the agent-specific requirements defined in this specification.

Provision	Agent-Specific Gap	T1	T2	T3
Art. 5	Board cannot characterise risk of unbounded action space	—	P	**[F]**
Art. 6(8)	Reference architecture is dynamic, not static	—	P	**[F]**
Art. 8(4)	ICT asset inventory changes at runtime	—	P	**[F]**
Art. 9	Inverted threat model: agent is actor, not target	—	**[F]**	**[F]**
Art. 10	Anomaly baseline undefined for variable behaviour	—	P	**[F]**
Art. 11	Mid-task interruption creates state corruption	—	P	**[F]**
Art. 12	Runtime state not capturable by periodic backup	—	P	**[F]**
Art. 17	Onset ambiguity for autonomous actions	—	P	**[F]**
Art. 19	Behavioural drift has no discrete onset for classification	—	P	**[F]**
Art. 24–25	Unbounded action space prevents exhaustive testing	—	P	**[F]**
Art. 26	Agent-specific threats absent from TLPT methodology	—	—	**[F]**
Art. 28	Dynamic tool connections create registration ambiguity	—	P	**[F]**
Art. 29–30	Tool inventory scales contractual burden linearly	—	P	**[F]**

*Table 2: DORA agent-specific gaps by autonomy tier. T1/T2/T3 = Tier 1/2/3. P = Partial gap. **[F]** = Full structural gap. Article 9 (Protection) is the only provision where the gap is full at Tier 2, reflecting the inverted threat model that applies as soon as the agent modifies external state autonomously.*

The pattern is consistent: Tier 1 agents present no material gaps. Tier 2 agents present partial gaps addressable through framework extensions. Tier 3 agents present full structural gaps across all thirteen provisions analysed. Article 9 is the sole exception to the tier gradient—its gap is full at Tier 2 because the inverted threat model (the agent as actor rather than target) activates the moment an agent modifies external state, regardless of whether it operates within a bounded envelope.

These gaps are not theoretical. Financial institutions deploying Tier 2 agents today operate with partial governance coverage. Institutions evaluating Tier 3 deployments face a regulatory environment that provides no guidance on how their most advanced AI systems should be governed. The remainder of this specification provides the framework that closes these gaps.

3 The AI Act's Human Oversight Paradox for Agents

3.1 The Structural Contradiction

Article 14 of Regulation (EU) 2024/1689 requires that high-risk AI systems be "designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use." The article specifies five capabilities that the oversight architecture must provide to the human overseer: the ability to fully understand the system's relevant capacities and limitations (Article 14(4)(a)); the ability to correctly interpret the system's output (14(4)(b)); the ability to decide not to use the system or to disregard, override, or reverse its output (14(4)(d)); and the ability to intervene in or interrupt the system through a "stop button" or similar procedure (14(4)(e)).

For non-agentic AI systems, these requirements are demanding but structurally coherent. A credit scoring model produces an output; the human reviews it; the human may override it. The output exists before any external action is taken, creating a temporal window for oversight.

For agentic AI systems, these requirements create a structural paradox. The defining characteristic of an agent is that it acts autonomously—it modifies external state without per-action

human approval. An agent that requires human approval for every action is not an agent; it is an assistive tool (Tier 1). The oversight requirement and the operational purpose of the agent are in direct tension: the more effective the oversight, the less autonomous the agent; the more autonomous the agent, the less effective the oversight.

This section analyses each Article 14 sub-requirement for agentic systems, identifies the specific point at which each requirement becomes structurally unachievable, and proposes an Agent Oversight Architecture that resolves the paradox through tier-appropriate intervention models.

3.2 Article 14(4)(a) — Understanding Capacities and Limitations

Understanding an Agent's Capacities and Limitations

Requirement: The human overseer must be able to "fully understand the relevant capacities and limitations" of the high-risk AI system and "be able to duly monitor its operation."

Non-Agentic Baseline

For a credit scoring model, capacities and limitations are characterisable at design time: the model processes defined inputs, produces scores within a known range, and has documented accuracy across defined populations. The human overseer can understand these through training and documentation.

Agent Paradox

An agent's capacities are not fixed at design time. They are a function of the agent's tool access (which may change), the model's interpretation of the goal (which varies per invocation), and the environmental context (which is runtime-dependent). The human overseer must understand not a static system but a dynamic capability space.

For a Tier 2 agent with a bounded action envelope, the overseer can understand the envelope—the set of approved actions—even if the agent's specific action sequence varies per invocation. The overseer understands *what the agent may do* without needing to predict *what the agent will do.*

For a Tier 3 agent, the capability space is defined by tool access and goal specification. The overseer cannot "fully understand" the system's capacities because the capacities emerge from the interaction between a non-deterministic model, a dynamic tool inventory, and a variable environmental context. The set of possible action sequences is combinatorially large and cannot be enumerated for human review.

The limitation side is equally problematic. The limitations of a non-agentic model are documentable: known failure modes, accuracy degradation conditions, population biases. The limitations of an agent include failure modes that are not discoverable at design time—novel combinations of tool outputs and environmental conditions that produce harmful actions without any individual component failing.

Resolution

For Tier 2 agents, Article 14(4)(a) is satisfied by documenting the action envelope, the escalation boundary, and the known failure modes of the agent within the envelope. The overseer understands the governance boundary, not the complete behaviour space.

For Tier 3 agents, Article 14(4)(a) cannot be satisfied through documentation alone. Compliance requires a **capability declaration** specifying: the tool inventory (what the agent can access), the goal specification framework (what objectives the agent can receive), the known interaction risks between tools and models, and an explicit acknowledgement of the residual capability space that cannot be fully characterised. This is honest compliance: the overseer understands what is knowable and is informed of what is not.

3.3 Article 14(4)(b) — Interpreting Output

Interpreting an Agent's Output

Requirement: The human overseer must be able to "correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available."

Non-Agentic Baseline

A credit scoring model's output is a score. Interpretation means understanding what the score represents, its confidence level, and the key features driving the score. Explainability tools (SHAP, LIME) provide feature attribution. The output is a single artifact that can be reviewed before action.

Agent Paradox

An agent's "output" is not a single artifact—it is an action sequence. A claims processing agent's output includes: documents retrieved, information extracted, policy terms interpreted, settlement amount calculated, and payment initiated. Each step is an output; the final action is the result of a chain of intermediate outputs.

"Correctly interpreting" this output requires understanding the entire action chain: why the agent retrieved those documents, how it interpreted the policy terms, why it calculated that settlement amount, and why it initiated payment rather than escalating. The causal chain may span dozens of intermediate reasoning steps across multiple tool interactions.

For Tier 2 agents, the action chain is bounded by the envelope and the intermediate steps can be logged for post-hoc review. The overseer cannot interpret the output in real time (the agent acts at machine speed) but can interpret it after the fact.

For Tier 3 agents, the action chain may be non-reproducible: the same input may produce a different action sequence on re-execution because the model's reasoning is non-deterministic and external system states may have changed. "Correctly interpreting" a non-reproducible action chain requires understanding the reasoning at the time of execution, not merely the logged artifacts.

Resolution

Agent architectures **SHALL** produce **action chain logs** that capture not merely the actions taken but the reasoning supporting each action: the goal decomposition, the tool selection rationale, the intermediate observations, and the decision criteria at each branch point. For Tier 3 agents, the logs must capture sufficient context to reconstruct the agent's reasoning state at each step, enabling post-hoc interpretation even when the action sequence is non-reproducible.

3.4 Article 14(4)(d) — Override and Reversal

Overriding or Reversing an Agent's Actions

Requirement: The human overseer must be able to "decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system."

Non-Agentic Baseline

For a credit scoring model, override means the human rejects the model's score and substitutes their own assessment. The model's output has not yet been acted upon; override occurs in the temporal window between output and action.

Agent Paradox

For an agent, the output *is* the action. By the time the human overseer is aware of the output, the agent has already: executed a trade (which has settled or is settling), sent a communication to a customer (which cannot be unsent), updated a database record (which downstream systems have already read), initiated a payment (which may have been received), or opened a regulatory filing (which has been submitted).

"Override" for non-agentic systems means preventing an action. "Override" for agentic systems means *reversing* an action that has already been executed. These are fundamentally different operations with fundamentally different feasibility profiles:

Reversible actions: database updates can be rolled back, draft communications can be recalled (within windows), pending payments can be cancelled. Override for these actions is technically feasible within time-bounded reversal windows.

Partially reversible actions: executed trades can be unwound but at market cost, customer communications can be followed by corrections but the original impression persists, claims payments can be adjusted but the customer has already received funds.

Irreversible actions: regulatory filings once submitted cannot be unfiled, real-time market impact cannot be undone, disclosed information cannot be un-disclosed, and certain contractual commitments cannot be withdrawn.

The temporal asymmetry is acute: the agent acts in milliseconds, the human reviews in minutes to hours. For a Tier 3 agent executing hundreds of actions per hour, the backlog of unreviewed actions grows continuously. The overseer is not overriding individual outputs—they are auditing a stream of completed actions and intervening when harm is detected, which may be well after the harm has occurred.

Resolution

The override requirement for agentic systems **SHALL** be implemented through three mechanisms calibrated to the reversibility profile of the agent's action space: (a) **pre-commitment gates** for irreversible actions—the agent must obtain human approval before executing any action classified as irreversible; (b) **time-bounded review windows** for partially reversible actions—the agent executes but the action enters a holding period during which the overseer can reverse; and (c) **continuous audit with correction authority** for reversible actions—the agent executes immediately and the overseer reviews asynchronously with full reversal capability.

3.5 Article 14(4)(e) — Interruption (The Stop Button)

Interrupting an Autonomous Agent

Requirement: The human overseer must be able to "intervene in the operation of the high-risk AI system or interrupt the system through a 'stop button' or a similar procedure that allows the system to come to a halt in a safe state."

Non-Agentic Baseline

For a non-agentic system, "halt in a safe state" means the system stops producing outputs. No external state has been modified; halting is clean.

Agent Paradox

For an agent mid-task, halting is not clean. The agent may be partway through a multi-step action sequence that has already modified external state. Halting produces one of three conditions:

Consistent halt: the agent has completed a logically atomic unit of work. External state is consistent. The halt is safe.

Inconsistent halt: the agent has completed some steps of a multi-step operation but not others. External state is partially updated. Downstream systems may read inconsistent data. The halt creates the CRT-7 (State Corruption) risk identified in the CRSA-1 EU Edition, extended to ART-5 (Cross-System State Propagation) for financial services agents.

Cascading halt: halting the agent does not halt the downstream effects of actions already taken. A trade executed before the halt continues to settle. A payment initiated before the halt continues to process. The "halt" stops the agent but not the consequences of the agent's prior actions.

Article 14(4)(e) requires the system to "come to a halt in a safe state." For a Tier 3 financial services agent, "safe state" must be defined not as the absence of agent activity but as the consistency of all systems the agent has touched. This requires transactional semantics across the agent's tool interactions—the ability to either complete or roll back an interrupted operation across multiple external systems. Most current agent frameworks do not provide this capability.

Resolution

The safe halt mechanism for agentic systems **SHALL** implement: (a) **immediate capability suspension**—the agent is prevented from initiating new actions; (b) **in-flight action assessment**—currently executing actions are classified as completable (non-state-modifying or within an atomic transaction) or abortable (state-modifying and not yet committed); (c) completable actions are allowed to finish; abortable actions are rolled back; (d) **state consistency verification**—all external systems the agent has modified during the current task are checked for consistency; (e) **inconsistency remediation**—detected inconsistencies are flagged for human resolution with full context of the agent's actions before and after the halt; and (f) **halt state logging**—the complete state of all agents, pending actions, and external system interactions is captured at the point of halt for forensic reconstruction.

3.6 The Agent Oversight Architecture

The preceding analysis demonstrates that Article 14 compliance for agentic systems cannot be achieved through a single oversight model. The structural paradox between autonomy and oversight is resolved not by eliminating either requirement but by matching the oversight model to the agent's autonomy tier and the reversibility profile of its action space.

This specification defines three intervention models, each providing a different resolution of the autonomy-oversight tension. Financial entities **SHALL** select the intervention model or combination of models appropriate to each agent's autonomy tier and action reversibility profile.

3.6.1 Intervention Model 1: Pre-Commitment Gates

Pre-Commitment Gate Model

The human overseer approves the agent's **action envelope**—the set of actions the agent may take autonomously—rather than approving each individual action. The agent operates freely within the approved envelope and must obtain explicit human approval before executing any action outside the envelope or any action classified as irreversible.

Applicable tiers: Tier 2 (primary model), Tier 3 (for irreversible actions only).

Oversight mechanism: The overseer defines and periodically reviews the action envelope. The agent enforces the envelope boundary at runtime. Actions outside the envelope are queued for human approval. The overseer's cognitive load is proportional to the rate of envelope violations and irreversible action requests, not to the total volume of agent actions.

Article 14 compliance:

- 14(4)(a): the overseer understands the envelope, not the complete behaviour space.
- 14(4)(b): actions within the envelope are interpretable through post-hoc log review; actions at the envelope boundary are reviewed before execution.
- 14(4)(d): override operates at the envelope level (the overseer can restrict the envelope) and at the action level (for out-of-envelope requests).
- 14(4)(e): halt suspends autonomous action within the envelope; the agent reverts to Tier 1 (assistive) mode.

3.6.2 Intervention Model 2: Real-Time Monitoring with Automatic Halt

Statistical Process Control Model

The agent operates autonomously while a monitoring system continuously evaluates the agent's action distribution against statistical baselines. When the monitoring system detects action patterns that deviate beyond defined control limits, the agent is automatically halted and the human overseer is notified for assessment.

Applicable tiers: Tier 2 (supplementary), Tier 3 (primary model for reversible actions).

Oversight mechanism: Statistical process control applied to agent behaviour. Control charts track action type distribution, tool utilisation patterns, escalation frequency, outcome distributions, and processing time profiles. The agent operates without per-action oversight but is subject to continuous statistical surveillance. Deviations trigger automatic halt and human assessment.

Article 14 compliance:

- 14(4)(a): the overseer understands the statistical baselines and control limits, which characterise "normal" agent behaviour.
- 14(4)(b): individual outputs are interpreted through post-hoc review; aggregate behaviour is interpreted through control chart analysis.
- 14(4)(d): override operates at the population level (the overseer adjusts control limits or suspends the agent) rather than the individual action level.
- 14(4)(e): automatic halt when control limits are breached provides a statistically grounded "stop button."

3.6.3 Intervention Model 3: Post-Action Audit with Reversal

Asynchronous Audit Model

The agent operates autonomously and all actions are executed immediately. A parallel audit process reviews completed actions against compliance criteria, business rules, and fairness metrics. Non-compliant actions are flagged for human review with full reversal capability for reversible actions and remediation procedures for irreversible actions.

Applicable tiers: Tier 3 (for reversible, low-stakes actions only).

Oversight mechanism: The audit process operates asynchronously at a cadence determined by the risk level: real-time sampling for high-volume actions, batch review for lower-volume actions. The human overseer reviews flagged actions and statistical summaries. Reversal is executed within the reversal window for each action type.

Article 14 compliance:

- 14(4)(a): the overseer understands the agent's operational profile through audit summaries and flagged action review.
- 14(4)(b): individual flagged outputs are fully interpretable through action chain logs; unflagged outputs are reviewed through statistical sampling.
- 14(4)(d): override operates through reversal of flagged actions and adjustment of the agent's parameters or scope.
- 14(4)(e): halt capability remains available but is not the primary control mechanism; the primary control is continuous audit with reversal.

3.7 Intervention Model Selection

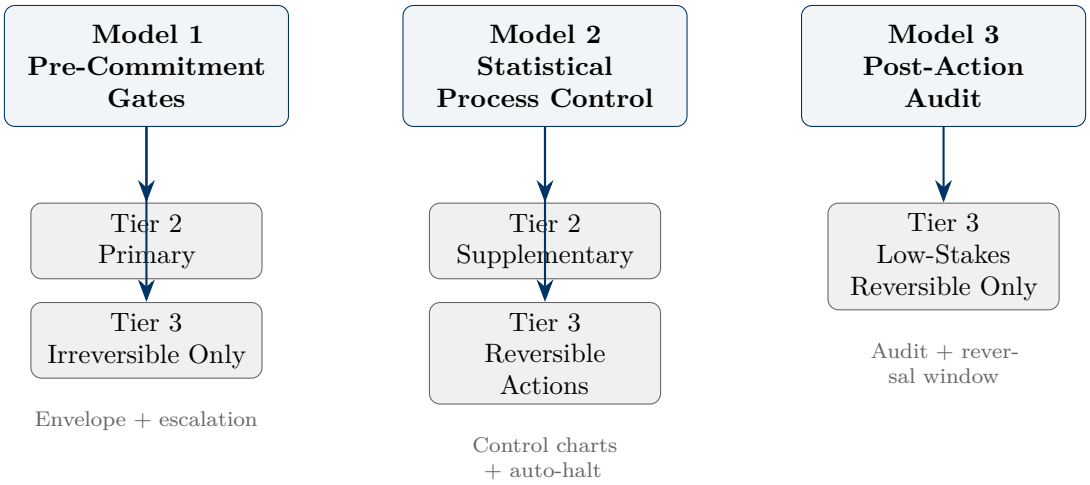

Figure 2: Agent Oversight Architecture. Three intervention models mapped to autonomy tiers and action reversibility profiles. Tier 3 agents require multiple concurrent models: pre-commitment gates for irreversible actions, statistical process control for reversible actions, and post-action audit for low-stakes reversible actions.

The critical design principle is that Tier 3 agents require **multiple concurrent intervention models**. No single model provides adequate Article 14 compliance for a fully autonomous agent. The agent's action space must be partitioned by reversibility: irreversible actions are governed by pre-commitment gates (Model 1), reversible actions by statistical process control (Model 2), and low-stakes reversible actions by post-action audit (Model 3). The combined architecture provides oversight at each level without requiring per-action human approval for the entire action space.

Appendix C provides the Agent Oversight Architecture Decision Matrix, which maps specific financial services action types to intervention models based on reversibility, regulatory impact, and customer harm potential.

3.8 Honest Framing: The Irreducible Oversight Gap

Honest Framing

The Agent Oversight Architecture resolves the structural paradox between Article 14 and autonomous agency for practical governance purposes. It does not eliminate the paradox.

The irreducible gap is this: for any Tier 3 agent executing reversible actions under Model 2 or Model 3, there exists a temporal window between the agent's action and the oversight system's detection of that action. During this window, the agent is acting without effective human oversight. The window may be milliseconds (for real-time statistical monitoring) or hours (for batch audit). No architecture can reduce this window to zero without converting the agent to Tier 1 (assistive mode), which eliminates the autonomy that justifies the agent's deployment.

The framework accepts this gap and manages it through three mechanisms: minimising the window through real-time monitoring, maximising reversal capability through transactional architecture, and bounding the maximum harm through action envelope constraints on irreversible actions. These mechanisms reduce the probability and severity of undetected harmful actions but do not eliminate the possibility.

Financial institutions deploying Tier 3 agents accept a non-zero probability of autonomous harmful actions occurring before detection. The question is not whether this residual risk exists—it does—but whether the institution's governance framework detects, bounds, and remediates harmful actions within tolerances approved by the management body under Article 5.

3.9 Additional AI Act Provisions for Agentic Systems

3.9.1 Article 9 — Risk Management for Agents

Article 9(2) requires identification and analysis of "known and reasonably foreseeable risks." For agentic systems, "reasonably foreseeable" extends beyond component-level risks to include emergent risks arising from the agent's autonomous interaction with its environment. The risk management system **SHALL** assess not only how the agent may fail (component failure) but how the agent may succeed at the wrong objective (goal misinterpretation), succeed through harmful means (means-end misalignment), or succeed in ways that produce unintended downstream effects (consequence blindness).

Article 9(4)'s requirement to assess "effects and possible interaction resulting from the combined application of the requirements" acquires particular force for agents. The interaction between logging granularity (Article 12) and agent processing speed creates tension: comprehensive action chain logging for a Tier 3 agent executing hundreds of actions per hour produces data volumes that may be impractical to store, review, or process. The risk management system must balance logging completeness against operational feasibility.

3.9.2 Article 13 — Transparency for Agents

Article 13(1) requires that high-risk AI systems be "designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret the system's output and use it appropriately." For agents, "interpreting the output" means interpreting the action sequence, not merely the final result. Transparency requires that the deployer can understand *why* the agent took each action in the sequence, which is a substantially higher transparency burden than explaining a single prediction or classification.

Article 13(3)(b) requires disclosure of the system's "capabilities and limitations." For agents, this disclosure must include: the capability envelope (tools and systems the agent can access), the action envelope (approved autonomous actions), the escalation boundary (conditions triggering human involvement), and the known behavioural limitations (conditions under which the agent's performance degrades or its actions become unpredictable).

3.9.3 Article 15 — Accuracy and Robustness for Agents

Article 15(2) requires declared accuracy metrics "in the instructions of use." For agents, accuracy cannot be expressed as a single metric. The specification **SHALL** include: task completion accuracy (proportion of tasks completed correctly), action fidelity (proportion of actions consistent with the stated objective), escalation accuracy (proportion of escalation decisions that were appropriate), and envelope compliance rate (proportion of actions within the approved envelope).

Article 15(3) requires resilience against "errors, faults or inconsistencies that may occur within the system or the environment within which the system operates." For agents, the "environment" is not a passive context—it is the set of external systems the agent actively modifies. Robustness must encompass resilience to unexpected environmental responses: tool failures, latency spikes, inconsistent external data, and concurrent modifications by other agents or human operators acting on the same external systems.

4 Agentic Risk Taxonomy for Financial Services

4.1 Purpose and Relationship to the Compositional Risk Taxonomy

The CRSA-1 EU Edition establishes the Compositional Risk Taxonomy (CRT-1 through CRT-8): eight categories of risk specific to composed AI systems, each arising from the interaction between components in a multi-model architecture. The CRT taxonomy is necessary but not sufficient for agentic systems. Agents introduce a class of risk that does not arise from composition alone but from *autonomous action within a dynamic environment*. A non-agentic composed system may exhibit cascade failure (CRT-1) or semantic drift (CRT-2) through component interaction, but it does not autonomously create new dependencies, cross regulatory boundaries, or propagate state changes across financial infrastructure—because it does not act autonomously.

This section establishes the **Agentic Risk Taxonomy (ART)**: six risk categories specific to autonomous AI agents operating in regulated financial services. Each ART category either extends a parent CRT category to the agentic context or defines a risk that has no compositional analogue—a risk that exists only because the system acts autonomously.

The ART taxonomy is designed to be applied alongside the CRT taxonomy, not to replace it. An agentic system is a composed system; it is subject to all applicable CRT risks *and* all applicable ART risks. The combined taxonomy provides the vocabulary for the architecture profiles in Section 5, the incident classification logic in Section 7, and the exit strategy framework in Section 8.

4.2 Taxonomy Architecture

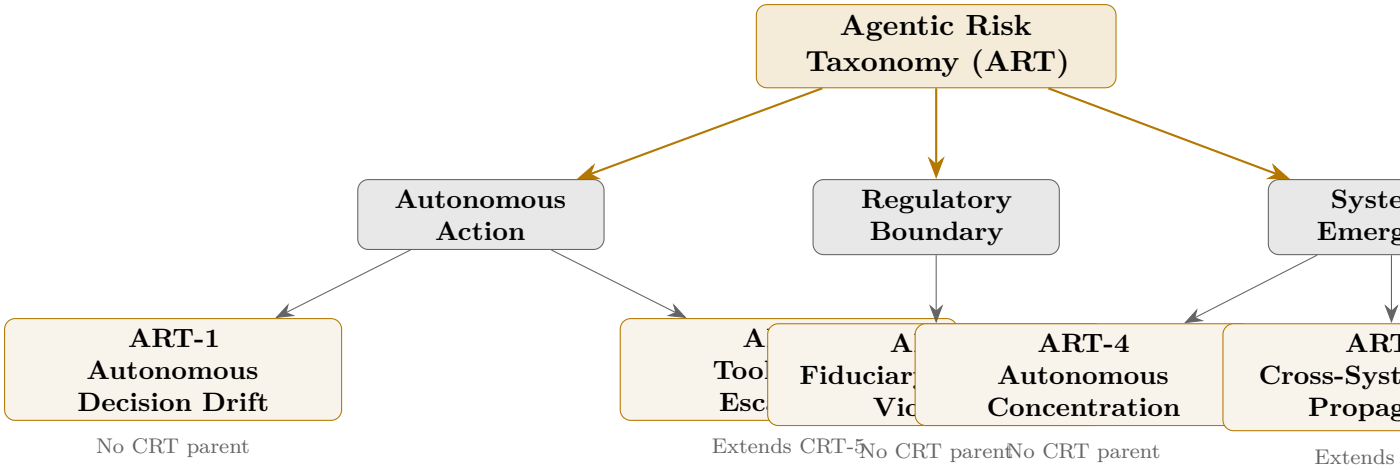

Figure 3: Agentic Risk Taxonomy. Six risk categories organised by origin: autonomous action (risks from the agent's self-directed behaviour), regulatory boundary (risks from the agent crossing regulatory perimeters), and systemic emergence (risks arising from the agent's interaction with financial infrastructure at scale). Parent CRT categories noted where applicable.

4.3 Risk Category Definitions

4.3.1 ART-1: Autonomous Decision Drift

Definition: Autonomous Decision Drift

Autonomous decision drift occurs when an agent's decision boundary shifts over time through accumulated context, self-directed prompt refinement, environmental feedback, or learned behavioural patterns, without a discrete model update event. Unlike model drift in non-agentic systems (where performance degrades because the data distribution shifts relative to the training distribution), autonomous decision drift arises because the agent's own history of actions and observations alters its subsequent behaviour. The agent that has processed 10,000 claims makes different decisions than the agent that has processed 100 claims, even if the underlying model has not changed, because its accumulated context, learned heuristics, and environmental feedback have shifted its effective decision boundary.

Origin: Autonomous action. No CRT parent—this risk category has no compositional analogue because it requires autonomous self-modification through experience.

Affected Regulations:

- DORA Article 10 — detection mechanisms must identify drift in the agent's decision distribution, not merely in the model's output distribution.
- DORA Article 19 — drift has no discrete onset, creating the cumulative incident classification problem identified in Section 2.
- AI Act Article 9 — risk management must assess the foreseeable risk that the agent's decisions will shift through accumulated experience.
- AI Act Article 15(2) — declared accuracy metrics may become stale as the agent's effective decision boundary shifts.
- AI Act Article 61 — post-market monitoring must detect autonomous decision drift as distinct from environmental change.
- SR 11-7 — ongoing monitoring must extend to the agent's behavioural trajectory, not merely its current performance.

Autonomy Tiers: Tier 2 (low severity—bounded envelope constrains drift impact). Tier 3 (high severity—unbounded action space allows drift to manifest in novel, unassessed behaviours).

Detection Methodology: Autonomous decision drift is detected through longitudinal analysis of the agent's action distributions. For a credit scoring agent, monitor the distribution of approval rates, score ranges, and adverse action reasons over rolling windows. When the distribution shifts beyond defined control limits (established during the agent's initial operating period), drift is indicated. The detection must distinguish between drift caused by changing input distributions (environmental shift—the population of applicants changed) and drift caused by the agent's accumulated context (autonomous drift—the agent's behaviour changed for the same input profile). Controlled test inputs (synthetic applicants with fixed profiles submitted periodically) provide the discrimination signal.

Mitigation Pattern: Autonomous decision drift is mitigated through context management: periodic resetting or pruning of the agent's accumulated context, version-controlled memory architectures that allow rollback to prior behavioural baselines, and mandatory recalibration intervals at which the agent's current behaviour is validated against the assessed baseline.

4.3.2 ART-2: Tool Chain Escalation

Definition: Tool Chain Escalation

Tool chain escalation occurs when an agent's dynamic tool selection produces a sequence of tool invocations whose combined capability exceeds any individually assessed tool's scope, granting the agent effective capabilities that were not anticipated in the conformity assessment or action envelope definition. The escalation is emergent: no single tool grants the excessive capability, but the chain of tool interactions produces it. A customer onboarding agent that queries a credit bureau, then accesses a transaction monitoring system, then invokes a communication tool has effectively assembled a surveillance and decision-making capability that none of those tools individually represents.

Origin: Autonomous action. Extends CRT-5 (Capability Drift) from the CRSA-1 EU Edition. CRT-5 addresses post-deployment capability changes through tool additions. ART-2 addresses capability emergence through *runtime tool composition*—the agent assembles capabilities dynamically that were not assessed as a combination.

Affected Regulations:

- DORA Article 8(4) — the combined capability produced by tool chain escalation may not appear in the ICT asset inventory because each tool is individually registered but the emergent combination is not.
- DORA Article 9 — protection against tool chain escalation requires capability-level access controls, not merely tool-level access controls.
- AI Act Article 3(23) — tool chain escalation may constitute a substantial modification if the emergent capability changes the system's compliance profile.
- AI Act Article 43(4) — if the combined capability was not predetermined, it triggers re-assessment.
- AI Act Article 51 — emergent combined capabilities may cross systemic risk thresholds without any individual tool change.

Autonomy Tiers: Tier 2 (medium—bounded envelope limits tool combinations). Tier 3 (critical—unbounded tool selection allows arbitrary capability assembly).

Detection Methodology: Tool chain escalation is detected through capability graph analysis: modelling each tool as a capability node and each tool chain as a capability path, then evaluating whether any observed path produces a combined capability exceeding the assessed scope. The analysis must run against the agent's runtime tool invocation logs, comparing observed tool chains against a pre-assessed library of acceptable combinations.

Mitigation Pattern: Tool chain escalation is mitigated through combination-level access controls: defining permitted tool chains (not merely permitted tools) in the action envelope, with runtime enforcement that evaluates each tool invocation in the context of the chain already executed during the current task. Chains exceeding the assessed combination scope are blocked and escalated.

4.3.3 ART-3: Fiduciary Boundary Violation

Definition: Fiduciary Boundary Violation

A fiduciary boundary violation occurs when an agent's autonomous actions cross a regulatory perimeter that the institution did not anticipate or design for, triggering obligations under a regulatory regime not addressed in the system's compliance framework. The violation is a consequence of the agent's goal-directed behaviour: the agent pursues its objective through the most effective action sequence available, without awareness of the regulatory boundaries between different financial activities. A customer service agent that autonomously offers a payment holiday on a credit product has crossed from customer service (limited-risk under the AI Act) into credit modification (potentially high-risk under Annex III point 5(b)). An advisory agent that autonomously generates a personalised investment recommendation has crossed from MiFID II information provision into personalised advice, triggering suitability assessment obligations.

Origin: Regulatory boundary. No CRT parent—this risk category has no compositional analogue because it requires autonomous action that crosses regulatory regimes, not merely interaction between components.

Affected Regulations:

- AI Act Article 6 — the system's risk classification may change mid-operation if the agent crosses from a non-high-risk activity into a high-risk activity.
- AI Act Article 25(1)(c) — an agent that autonomously modifies its effective intended purpose triggers provider reclassification.
- MiFID II Article 24–25 — an agent crossing from information provision to personalised advice triggers suitability assessment obligations.
- Consumer Credit Directive — an agent that autonomously modifies credit terms triggers pre-contractual information and creditworthiness reassessment obligations.
- GDPR Article 22 — an agent that autonomously makes a decision producing legal effects on a natural person triggers the prohibition on solely automated decision-making without appropriate safeguards.
- AML Directive — an agent that accesses customer identity information for one purpose and uses it for another may violate purpose limitation requirements.

Autonomy Tiers: Tier 2 (low—bounded envelope constrains regulatory boundary crossings if the envelope is regulation-aware). Tier 3 (critical—goal-directed behaviour inherently seeks the most effective action regardless of regulatory perimeter).

Detection Methodology: Fiduciary boundary violation is detected through regulatory perimeter classification of the agent's actions. Each action type in the agent's capability inventory is pre-classified by the regulatory regime it engages (AI Act risk level, MiFID II activity category, consumer credit obligation, AML obligation). The agent's runtime action sequence is evaluated against this classification to detect transitions between regulatory regimes during a single task execution.

Mitigation Pattern: Fiduciary boundary violation is mitigated through regulation-aware action envelope design: the action envelope is partitioned by regulatory regime, and transitions between regimes require explicit escalation to human oversight or pre-authorised cross-regime action

protocols that ensure the institution's compliance framework is activated for the destination regime before the agent acts within it.

4.3.4 ART-4: Autonomous Concentration

Definition: Autonomous Concentration

Autonomous concentration occurs when multiple agents operating independently across a financial institution converge on the same strategy, the same tool providers, or the same market positions without any single human decision directing the convergence. Each agent independently optimises its objective and independently arrives at a similar solution, creating institutional concentration risk that emerges from the aggregate behaviour of autonomous systems rather than from deliberate strategic choice. A bank deploying separate trading agents for equities, fixed income, and derivatives may find all three agents independently building exposure to the same macroeconomic factor, creating a concentrated risk position that no individual risk manager authorised.

Origin: Systemic emergence. No CRT parent—this risk category requires multiple autonomous agents whose independent optimisation produces emergent concentration.

Affected Regulations:

- DORA Article 28(4)(c) — concentration risk assessment must detect agent-driven concentration, not merely human-directed vendor selection.
- DORA Article 6(9) — the multi-vendor strategy must account for autonomous vendor selection by agents.
- AI Act Article 9(4) — the "effects and possible interaction" between agents must be assessed as part of risk management.
- CRR/CRD — large exposure limits and concentration risk frameworks must capture agent-created exposures.
- FSB AI financial stability framework — systemic concentration through AI represents an emerging macroprudential risk.

Autonomy Tiers: Tier 2 (medium—bounded envelopes limit but do not eliminate convergence). Tier 3 (critical—unbounded optimisation maximises convergence probability).

Detection Methodology: Autonomous concentration is detected through cross-agent correlation analysis: monitoring the portfolio of all active agents within the institution for position correlation, vendor dependency correlation, strategy correlation, and tool utilisation correlation. When cross-agent correlation exceeds defined thresholds, concentration is indicated. The analysis must operate at the institutional level, not the individual agent level, because the concentration is an emergent property of the agent population.

Mitigation Pattern: Autonomous concentration is mitigated through institutional-level agent coordination: a meta-monitoring layer that observes all agents' positions, strategies, and tool usage, and imposes diversity constraints when correlation thresholds are approached. Individual agents may be required to randomise certain decisions (tool selection, timing, routing) to prevent emergent convergence.

4.3.5 ART-5: Cross-System State Propagation

Definition: Cross-System State Propagation

Cross-system state propagation occurs when an agent modifies state in one external system, and that modification triggers cascading effects in other systems that the agent does not observe, monitor, or control. Unlike CRT-7 (State Corruption), which addresses inconsistent state from interrupted operations, ART-5 addresses *consistent but unintended* state propagation: the agent's action succeeds, the external system processes it correctly, and the downstream cascade proceeds as designed—but the agent did not anticipate or intend the downstream effects. A claims processing agent that initiates a large payment triggers the bank's AML transaction monitoring system, which flags the payment, which freezes the customer's account, which prevents the customer from accessing funds—a cascade the claims agent neither intended nor observes.

Origin: Systemic emergence. Extends CRT-7 (State Corruption) from the CRSA-1 EU Edition. CRT-7 addresses inconsistent state from component failure. ART-5 addresses consistent but unintended propagation from successful agent action across interconnected financial systems.

Affected Regulations:

- DORA Article 11 — response and recovery must address cascading effects across systems the agent does not directly control.
- DORA Article 12 — logging must capture not only the agent's direct actions but the downstream effects in connected systems.
- AI Act Article 9 — risk management must assess the foreseeable downstream effects of the agent's actions in interconnected financial infrastructure.
- AI Act Article 14(4)(e) — the "stop button" halts the agent but does not halt downstream propagation already initiated.
- AI Act Article 15(3) — robustness must encompass resilience to unintended cascading effects from the agent's own actions.

Autonomy Tiers: Tier 2 (medium—bounded actions have assessable downstream effects). Tier 3 (high—dynamic action sequences create unpredictable propagation paths).

Detection Methodology: Cross-system state propagation is detected through downstream impact monitoring: instrumenting the external systems the agent modifies to report state changes and cascading triggers back to the agent's monitoring infrastructure. The monitoring must extend beyond the agent's direct interactions to include secondary and tertiary effects in connected systems.

Mitigation Pattern: Cross-system state propagation is mitigated through impact pre-assessment: before executing state-modifying actions, the agent (or its oversight infrastructure) evaluates the known downstream dependencies of the target system and assesses whether the intended action may trigger cascading effects beyond the agent's scope. Actions with high propagation risk are escalated to human oversight or executed with monitoring of downstream systems active.

4.3.6 ART-6: Accountability Void

Definition: Accountability Void

An accountability void occurs when, in a multi-agent delegation chain, no single agent's logs capture the complete causal chain from the initial task input to the adverse outcome. Each agent logs its own actions and reasoning, but the causal connection between the coordinating agent's delegation decision, the specialist agent's execution, the tool's response, and the resulting harm is distributed across multiple log systems with no unified causal reconstruction capability. The accountability void is not a logging failure—each agent logs correctly—but an architectural gap: the system lacks a cross-agent causal model that connects distributed logs into a single accountable narrative. When a regulator or affected person asks "why did the system do this?", no single log provides the answer, and reconstructing the answer requires cross-referencing multiple agent logs, tool interaction records, and external system states—a forensic exercise that may be infeasible within incident reporting timelines.

Origin: Systemic emergence. Extends CRT-8 (Compositional Opacity) from the CRSA-1 EU Edition. CRT-8 addresses the difficulty of explaining decisions produced by composed systems. ART-6 addresses the deeper problem of *attributing accountability* when the causal chain is distributed across autonomous agents.

Affected Regulations:

- DORA Article 17 — incident investigation requires reconstructing the causal chain, which the accountability void impedes.
- DORA Article 19 — the 4-hour initial notification requires rapid attribution, which distributed logs frustrate.
- AI Act Article 12 — logging must enable identification of risk situations, which requires cross-agent causal reconstruction.
- AI Act Article 13 — transparency requires the deployer to interpret the system's output, which requires understanding the delegation chain.
- AI Act Article 86 — the right to explanation for affected persons requires attributing the decision to specific causal factors, which the accountability void prevents.
- GDPR Article 22 — the right not to be subject to solely automated decision-making requires the institution to explain the decision logic, which distributed agent chains obscure.

Autonomy Tiers: Tier 2 (low—single-agent systems with human escalation maintain clear accountability). Tier 3 (critical—multi-agent delegation chains distribute accountability across autonomous actors).

Detection Methodology: The accountability void is detected through causal reconstruction testing: periodically selecting completed agent tasks and attempting to reconstruct the complete causal chain from input to outcome using only the available logs. When reconstruction fails (missing causal links, unexplained delegation decisions, tool interactions without recorded context), the accountability void is confirmed. The test must be performed within the incident reporting timeline (4 hours for DORA) to verify that reconstruction is feasible within regulatory deadlines.

Mitigation Pattern: The accountability void is mitigated through distributed causal logging: a unified correlation framework that assigns a single task identifier to the entire multi-agent execution, propagates it through all delegation chains and tool interactions, and produces a reconstructable causal graph linking every agent action, delegation decision, and tool response to the originating task. The causal graph must be queryable within incident reporting timelines.

4.4 Cross-Reference: ART Categories to Regulatory Provisions

ART Category	DORA Provisions	AI Act Provisions	Other Regulation	CRT Parent
ART-1 Decision Drift	Art. 10, 19	Art. 9, 15(2), 61	SR 11-7	None
ART-2 Tool Chain Escalation	Art. 8(4), 9	Art. 3(23), 43(4), 51	—	CRT-5
ART-3 Fiduciary Boundary	—	Art. 6, 25(1)(c)	MiFID II, CCD, GDPR Art. 22, AMLD	None
ART-4 Autonomous Concentration	Art. 6(9), 28(4)(c)	Art. 9(4)	CRR/CRD, FSB	None
ART-5 State Propagation	Art. 11, 12	Art. 9, 14(4)(e), 15(3)	—	CRT-7
ART-6 Accountability Void	Art. 17, 19	Art. 12, 13, 86	GDPR Art. 22	CRT-8

Table 3: Agentic Risk Taxonomy cross-reference. Each ART category mapped to primary DORA provisions, AI Act provisions, other applicable financial regulation, and parent CRT category where applicable. Three ART categories (ART-1, ART-3, ART-4) have no CRT parent—they are uniquely agentic risks with no compositional analogue.

4.5 Taxonomy Completeness and Extensibility

This taxonomy is designed to be exhaustive with respect to the agentic deployment patterns identified in regulated financial services as of March 2026. The six categories are organised along three origin dimensions (autonomous action, regulatory boundary, systemic emergence) that provide a complete classification of the sources of agent-specific risk.

The taxonomy is structured for extensibility. As agentic architectures evolve—particularly as multi-agent systems become more prevalent and agents acquire longer-horizon planning capabilities—new risk categories may emerge. Additional categories **SHALL** be assigned sequential ART identifiers (ART-7 and beyond), classified by origin dimension, and mapped to the regulatory provisions they affect.

Three risk categories have no CRT parent: ART-1 (Autonomous Decision Drift), ART-3 (Fiduciary Boundary Violation), and ART-4 (Autonomous Concentration). These represent genuinely novel risk categories that exist only in the agentic context. Three categories extend CRT parents: ART-2 extends CRT-5, ART-5 extends CRT-7, ART-6 extends CRT-8. For systems where both the CRT parent and the ART extension apply, both **SHALL** be assessed—the CRT category captures the compositional risk component and the ART category captures the agent-specific amplification.

The ART identifiers are referenced throughout the remainder of this specification. Section 5 maps each architecture profile to its applicable ART categories. Section 7 structures incident classification around the ART framework. Appendix A incorporates ART categories into the Agent Capability Declaration Template.

5 Financial Architecture Profiles for Agentic Systems

5.1 Purpose and Structure

The DORA gap analysis in Section 2 establishes where existing regulation fails for agents. The oversight paradox in Section 3 establishes how to govern agents. The ART taxonomy in Section 4 establishes what can go wrong. This section establishes *how specific financial agentic architectures must be designed, documented, monitored, and assessed* to satisfy those requirements and manage those risks.

Each profile addresses a dominant agentic deployment pattern in financial services. For each architecture, the profile provides:

1. An architecture reference diagram identifying agent components, tool connections, delegation chains, and compliance-critical boundaries.
2. The applicable ART and CRT risk categories ranked by severity for that architecture.
3. The autonomy tier classification and the applicable intervention model(s) from the Agent Oversight Architecture.
4. DORA-specific compliance requirements.
5. AI Act classification analysis and compliance requirements.
6. The safe halt specification for that architecture.

5.2 Profile A: Autonomous Trade Execution Agent

5.2.1 Architecture Reference

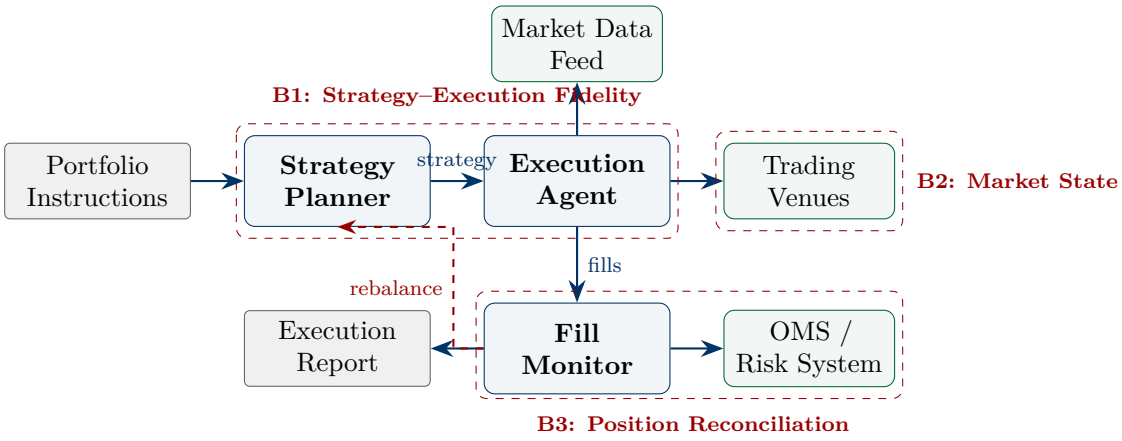

Figure 4: Autonomous trade execution agent. Three compliance-critical boundaries: B1 (fidelity between strategy planner intent and execution agent action), B2 (market state—the agent modifies market conditions through its own trades), B3 (position reconciliation between agent-reported fills and OMS/risk system state). Dashed red line indicates feedback loop from fill monitoring to strategy replanning.

5.2.2 Autonomy Tier and Oversight Model

Autonomy classification: Tier 3 (Fully Autonomous). The agent receives portfolio-level instructions and independently determines execution strategy, venue selection, order timing, order sizing, and rebalancing decisions. External state modification (trade execution) is the agent's core function.

Intervention models:

- **Model 1 (Pre-Commitment Gates):** applied to position limits, single-order size thresholds, venue restrictions, and instrument scope. The agent cannot exceed pre-approved exposure limits without human authorisation.

- **Model 2 (Statistical Process Control):** applied to execution quality metrics (VWAP deviation, market impact, fill rate), venue concentration, and rebalancing frequency. Control limits trigger automatic halt when execution behaviour deviates from assessed baselines.
- **Model 3 (Post-Action Audit):** applied to individual order routing decisions and venue selection within approved parameters. Audit operates on T+0 basis with full reversal capability through offsetting trades (at market cost).

5.2.3 Applicable Risk Categories

Risk Category	Severity	Trade Execution Manifestation
ART-4 Autonomous Concentration	Critical	Multiple trading agents across desks independently converge on the same position, creating institutional exposure no individual desk authorised.
ART-5 State Propagation	Critical	Agent executes a large order; fill triggers downstream margin calls, collateral movements, and counterparty exposure recalculations that the agent does not observe.
ART-1 Decision Drift	High	Execution strategy evolves through accumulated market observations: the agent that has traded through a volatile period develops different venue preferences and timing heuristics than initially assessed.
ART-2 Tool Chain Escalation	High	Agent combines market data, news sentiment analysis, and venue analytics to construct an execution strategy that constitutes de facto portfolio management—exceeding the scope of execution-only instructions.
ART-3 Fiduciary Boundary	High	Agent optimises execution in a way that constitutes market-making activity (providing liquidity through passive orders), crossing from agency execution into principal trading territory.
ART-6 Accountability Void	Medium	Strategy planner delegates to execution agent which interacts with multiple venues; when best execution obligations are questioned, reconstructing the complete decision chain across agents and venues is forensically complex.
CRT-1 Cascade Failure	Medium	Strategy planner misinterprets portfolio instructions; execution agent faithfully executes the wrong strategy.

5.2.4 Regulatory Classification

AI Act: Trade execution algorithms are not listed in Annex III. The system is not classified as high-risk under the AI Act unless the portfolio allocation decision feeding the agent was itself produced by a high-risk AI system (in which case the execution agent is part of the composed high-risk system under Article 25(1)). If the portfolio instructions originate from human portfolio managers, the execution agent operates as a limited-risk or non-classified system.

MiFID II: Article 17 applies directly. Algorithmic trading systems must have effective systems and risk controls, with annual self-assessment reported to the competent authority. Article 27 best execution obligations apply to every order the agent routes.

DORA: Fully applicable. The trading system supports a critical business function. All DORA ICT risk management, incident reporting, resilience testing, and third-party risk provisions apply at the highest criticality level.

5.2.5 DORA Compliance Requirements

Article 8(4) — ICT Asset Identification. The static capability inventory **SHALL** include: the strategy planner model (provider, version, API endpoint), all connected market data feeds (providers, data types, latency specifications), all connected trading venues (venue identifiers, connectivity type, protocol versions), the OMS/risk system integration (read/write permissions, reconciliation frequency), and the fill monitoring infrastructure. The dynamic runtime inventory **SHALL** log, per execution session: which venues were accessed, which data feeds were queried, and which rebalancing iterations were executed.

Article 10 — Detection. Three-tier monitoring **SHALL** be implemented. *Envelope compliance*: position limits, order size limits, venue restrictions, instrument scope—immediate alert on breach. *Behavioural distribution*: VWAP deviation distribution, venue utilisation distribution, rebalancing frequency, average market impact—control charts with defined limits. *Outcome quality*: execution cost versus benchmark, slippage rate, fill rate, best execution compliance—continuous tracking against MiFID II Article 27 requirements.

Article 19 — Incident Classification. The following agent-specific events **SHALL** be assessed against the classification criteria: single-order loss exceeding a defined threshold (action-as-incident assessment), cumulative execution quality degradation over a defined window (cumulative onset assessment), unexpected venue concentration exceeding the assessed baseline (autonomous concentration indicator), and position reconciliation failure between the agent's reported fills and the OMS (state propagation indicator).

Article 24–26 — Resilience Testing. The testing programme **SHALL** include: market data feed failure simulation (the agent loses access to its primary data source mid-execution), venue unavailability (the agent's preferred venue becomes unreachable), latency spike injection (market data arrives with variable delay), adversarial market data (manipulated price feeds testing the agent's data validation), and safe halt during active execution (verifying that open orders are cancelled and positions are reconciled upon halt).

5.2.6 Safe Halt Specification

The safe halt mechanism for the trade execution agent **SHALL** execute the following sequence:

1. **Immediate capability suspension:** the strategy planner is prevented from generating new execution instructions.
2. **Open order cancellation:** all pending orders across all venues are cancelled. Cancellation confirmation is required from each venue before proceeding.
3. **In-flight order assessment:** orders that have been partially filled are assessed. The remaining unfilled portion is cancelled. The filled portion is recorded.
4. **Position reconciliation:** the agent's internal position state is reconciled against the OMS and each venue's confirmed fills. Discrepancies are flagged for human resolution.
5. **Risk system notification:** the OMS and risk system are updated with the agent's final position state, enabling downstream risk calculations to reflect the halted state.
6. **Halt state log:** the complete state at halt is logged: all open orders at the time of halt, cancellation confirmations, partial fill details, position reconciliation results, and the reason for halt.

CRSA-1 EU Compliance Series — Forthcoming

Specific execution quality threshold parameters, control chart specifications for trading agent behavioural monitoring, and MiFID II Article 27 best execution compliance metrics for autonomous execution agents will be provided in implementation guidance accompanying the CRSA-1 Financial Services Edition.

5.3 Profile B: AML Investigation Agent

5.3.1 Architecture Reference

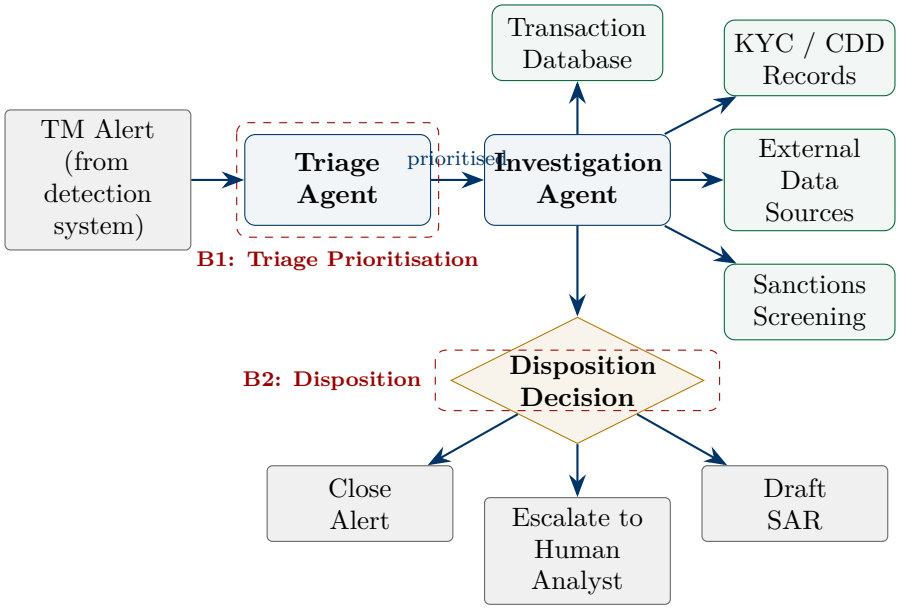

Figure 5: AML investigation agent. Two compliance-critical boundaries: B1 (triage prioritisation—which alerts the agent investigates first affects detection timing), B2 (disposition decision—the agent autonomously determines whether to close, escalate, or draft a SAR). The investigation agent accesses multiple data sources dynamically based on alert characteristics.

5.3.2 Autonomy Tier and Oversight Model

Autonomy classification: The AML investigation agent spans Tier 2 and Tier 3 depending on the disposition pathway. Alert triage and evidence gathering are Tier 2 (supervised autonomous within a defined investigation protocol). Alert closure and SAR drafting are Tier 3 functions if executed without per-action human approval.

Intervention models:

- **Model 1 (Pre-Commitment Gates):** applied to disposition decisions. Alert closure **SHALL** require human confirmation for alerts above a defined risk threshold. SAR filing **SHALL** require human review in all cases—the regulatory obligation to file a SAR is the institution's, not the agent's, and autonomous SAR filing without human review creates accountability risk.
- **Model 2 (Statistical Process Control):** applied to triage prioritisation patterns, investigation depth (number of data sources accessed per alert), closure rate, escalation rate, and average investigation duration. Control limits detect behavioural drift in investigation patterns.
- **Model 3 (Post-Action Audit):** applied to evidence gathering actions within an investigation. The agent's data source selections and query patterns are audited for completeness and relevance.

5.3.3 Applicable Risk Categories

Risk Category	Severity	AML Investigation Manifestation
ART-3 Fiduciary Boundary	Critical	Investigation agent accessing customer financial data for AML purposes autonomously uses the same data to assess creditworthiness or insurance risk, crossing from AML (exempt from AI Act) into high-risk territory (Annex III point 5(b)).
ART-1 Decision Drift	Critical	Investigation agent's closure threshold shifts over time through accumulated case exposure. After processing thousands of low-risk alerts, the agent's effective threshold for "suspicious" recalibrates upward, closing alerts that earlier iterations would have escalated.
ART-6 Accountability Void	High	Triage agent prioritises, investigation agent gathers evidence and decides disposition. When a closed alert later proves to be genuine money laundering, reconstructing why the triage agent deprioritised it and why the investigation agent closed it requires cross-agent causal reconstruction under regulatory examination pressure.
ART-2 Tool Chain Escalation	High	Agent combines transaction data, KYC records, external data sources, and sanctions screening into a comprehensive customer profile that constitutes mass surveillance capability—exceeding the scope of the individual alert investigation.
CRT-6 Context Poisoning	High	External data sources consulted during investigation contain manipulated information designed to cause the agent to close legitimate alerts or escalate false positives.
ART-5 State Propagation	Medium	Agent files a SAR; the filing triggers an account freeze by the compliance system, which prevents the customer from transacting, which triggers customer complaints—a cascade the investigation agent neither intended nor observes.

5.3.4 Regulatory Classification

AI Act: The AML investigation agent operates at the boundary of the financial fraud detection exemption. Annex III point 5(b) exempts "AI systems used for the purpose of detecting financial fraud." The Commission's May 2025 clarification confirmed that fraud detection and AML transaction monitoring AI are exempt from high-risk classification. However, the exemption applies to the *purpose* of fraud detection—if the agent autonomously repurposes investigation data for creditworthiness assessment or insurance pricing (ART-3), the exemption is lost for that action and the system is reclassified as high-risk for that function.

AML Directive: The institution bears ultimate responsibility for suspicious transaction reporting. The agent may assist but cannot relieve the institution of its obligation to ensure adequate investigation and timely filing. Autonomous alert closure without human review creates regulatory risk: if a closed alert later proves to involve money laundering, the institution must demonstrate that the closure decision met the standard of care—a demonstration that is substantially harder when the decision was made by an autonomous agent.

DORA: Fully applicable. AML transaction monitoring supports a critical business function (regulatory compliance). All DORA provisions apply at the highest criticality level.

GDPR: Article 22 is engaged if the agent's disposition decision produces legal effects on the customer (account freeze, SAR filing). The institution must ensure appropriate safeguards including the right to obtain human intervention, express a point of view, and contest the decision.

5.3.5 DORA Compliance Requirements

Article 8(4) — ICT Asset Identification. The static capability inventory **SHALL** include: the triage agent model, the investigation agent model, all connected data sources (transaction database, KYC/CDD records, external data providers, sanctions screening services), and the SAR filing system. The dynamic runtime inventory **SHALL** log, per investigation: which data sources were queried, what queries were executed, and what disposition was reached.

Article 10 — Detection. Three-tier monitoring **SHALL** track: *Envelope compliance*: the agent does not access data sources outside the approved investigation protocol, does not query customer data beyond the scope of the specific alert. *Behavioural distribution*: closure rate, escalation rate, SAR drafting rate, average investigation depth, and data source utilisation patterns—all monitored for drift. *Outcome quality*: false negative rate (alerts closed by the agent that are subsequently identified as genuine suspicious activity through other channels), false positive escalation rate, and SAR quality metrics (completeness, accuracy of narrative).

Article 17–19 — Incident Classification. The following agent-specific events **SHALL** be assessed: a closed alert subsequently confirmed as genuine money laundering (action-as-incident—the onset is the closure decision, detected retrospectively), systematic under-investigation of alerts from a specific customer segment (cumulative onset—detected through statistical analysis of investigation depth by segment), and manipulation of external data sources causing incorrect dispositions (CRT-6 context poisoning—onset at the point of data compromise).

5.3.6 Safe Halt Specification

1. **Immediate capability suspension:** both triage and investigation agents are prevented from processing new alerts or continuing active investigations.
2. **In-progress investigation assessment:** investigations in progress are catalogued with their current state (evidence gathered, data sources consulted, preliminary findings).
3. **Alert queue preservation:** all unprocessed alerts remain in the queue with their original priority and timestamp, available for human analyst processing.
4. **Disposition hold:** any pending disposition decisions (close, escalate, SAR) that have not been executed are frozen for human review.
5. **Regulatory timeline assessment:** for any alerts approaching SAR filing deadlines, the halt state log flags these for immediate human attention to prevent regulatory deadline breach.
6. **Halt state log:** complete state captured including all active investigations, evidence gathered, pending dispositions, and alert queue status.

The AML-specific safe halt concern is **regulatory deadline preservation**. AML regulations impose filing deadlines for SARs (typically 30 days from determination of suspicious activity in most Member State implementations). If the agent is halted mid- investigation, the institution must ensure that time-sensitive alerts are immediately transferred to human analysts with full context to prevent deadline breach. The halt state log **SHALL** include a regulatory deadline flag for each active investigation.

5.4 Profile C: Claims Processing Agent

5.4.1 Architecture Reference

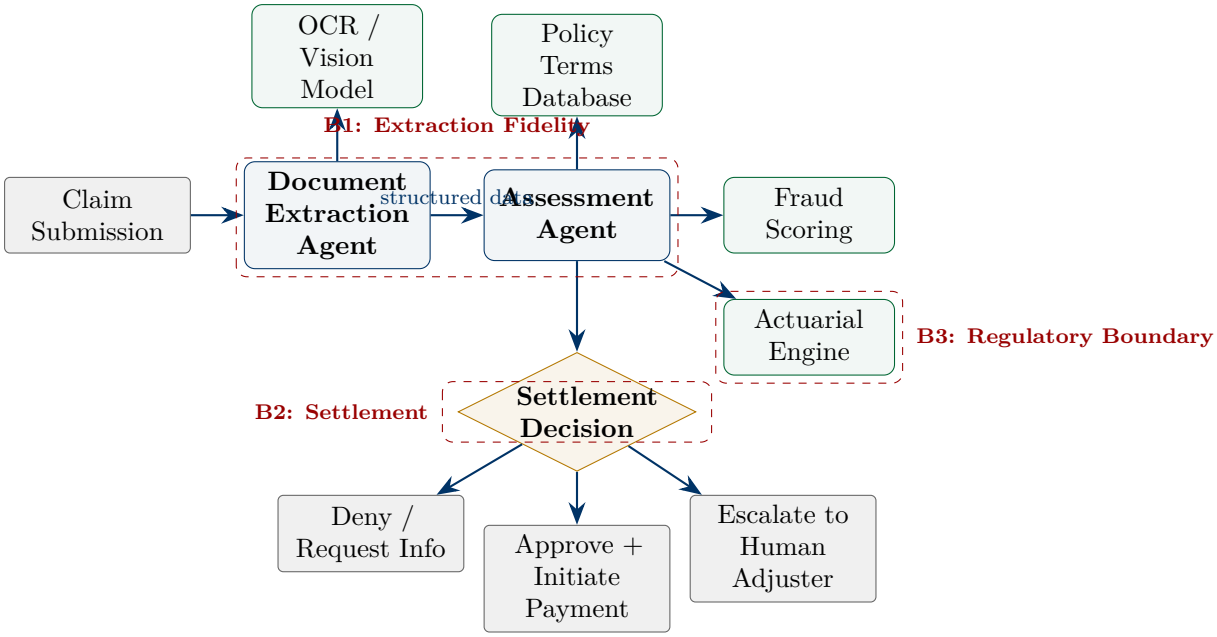

Figure 6: Claims processing agent. Three compliance-critical boundaries: B1 (fidelity between document extraction and structured data used for assessment—OCR/vision model errors cascade into incorrect settlement calculations), B2 (settlement decision—the agent autonomously determines claim outcome and payment amount), B3 (regulatory boundary—the actuarial engine connection means the agent may cross from claims processing into pricing/risk assessment territory, engaging Annex III point 5(c) for life and health insurance).

5.4.2 Autonomy Tier and Oversight Model

Autonomy classification: Tier 2 for routine claims within policy limits (supervised autonomous with defined settlement parameters). Tier 3 for complex claims requiring policy interpretation, multi-document synthesis, or settlement amounts approaching policy limits.

Intervention models:

- **Model 1 (Pre-Commitment Gates):** applied to settlement amount thresholds (claims above a defined value require human approval), claim denial decisions (all denials reviewed by human adjuster before communication to policyholder), and any claim involving bodily injury or fatality.
- **Model 2 (Statistical Process Control):** applied to approval rate, average settlement amount, denial rate, fraud referral rate, and processing time—all monitored for drift across claim categories, policy types, and claimant demographics.
- **Model 3 (Post-Action Audit):** applied to approved claims below the pre-commitment threshold. Random sampling audit of settlement calculations, policy term interpretation accuracy, and payment initiation correctness.

5.4.3 Applicable Risk Categories

Risk Category	Severity	Claims Processing Manifestation
ART-3 Fiduciary Boundary	Critical	Agent accesses the actuarial engine to verify coverage terms and inadvertently performs risk re-assessment, crossing from claims processing (not in Annex III) into risk assessment and pricing for life/health insurance (Annex III point 5(c)). The system's AI Act classification changes mid-operation from non-high-risk to high-risk.
ART-5 State Propagation	Critical	Agent initiates a large settlement payment; the payment triggers the bank's AML transaction monitoring, which flags it as unusual activity, which freezes the policyholder's account pending investigation—a cascade the claims agent neither intended nor observes.
ART-1 Decision Drift	High	Agent's settlement calibration shifts through accumulated claim exposure. After processing thousands of minor claims, the agent's implicit baseline for "reasonable settlement" recalibrates, producing systematically higher or lower settlements than the initial assessed baseline.
CRT-1 Cascade Failure	High	OCR/vision model misreads a document (wrong date, wrong amount, wrong policy number); the assessment agent processes the incorrect data as valid, producing a settlement based on the wrong claim facts. Neither component fails—the extraction is confident and the assessment is correct given its input.
CRT-3 Aggregation Bias	Medium	If the fraud scoring model and the assessment agent independently exhibit demographic bias, the combined effect on claim outcomes may produce discriminatory settlement patterns that neither model exhibits individually.
ART-6 Accountability Void	Medium	Document extraction agent feeds assessment agent which queries fraud scoring and actuarial engine. When a policyholder disputes a denied claim, reconstructing why the agent denied—which extracted data element, which policy term interpretation, which fraud score—requires cross-agent causal reconstruction.

5.4.4 Regulatory Classification

AI Act: The claims processing agent operates at a critical Annex III boundary. Claims adjudication—determining whether a claim is valid and calculating the settlement amount—is not itself listed in Annex III. However, Annex III point 5(c) covers "AI systems intended to be used for risk assessment and pricing in relation to natural persons in the case of life and health insurance." If the agent's assessment process involves recalculating risk (through the actuarial engine connection at Boundary B3), the agent has crossed into high-risk territory.

The classification depends on the agent's interaction with the actuarial engine. If the agent queries the engine in read-only mode to verify coverage terms and policy limits ("what does this policy cover?"), the interaction is informational and does not constitute risk assessment. If the agent invokes the engine to calculate risk-adjusted settlement amounts or to reassess the claim against the policyholder's current risk profile ("should this claim be settled at face value given updated risk factors?"), the agent is performing risk assessment and the system is reclassified as high-risk for that function.

Solvency II: The claims processing function falls within the Solvency II governance framework. The insurer's ORSA (Own Risk and Solvency Assessment) must account for operational risk from autonomous claims processing, including the risk of systematic over- or under-settlement.

Insurance Distribution Directive (IDD): If the agent communicates directly with the policyholder (providing claim status, explaining denial reasons, offering settlement terms), the communication may constitute an insurance distribution activity requiring compliance with IDD conduct of business requirements.

DORA: Fully applicable. Claims processing supports a critical business function. The agent's connections to external data sources and payment systems create ICT third-party dependencies requiring Article 28–30 compliance.

GDPR: Article 22 is directly engaged. A claim denial or settlement determination by an autonomous agent produces a decision "based solely on automated processing" that "produces legal effects" on the policyholder. The insurer must ensure appropriate safeguards, meaningful information about the logic involved, and the right to obtain human intervention.

5.4.5 DORA Compliance Requirements

Article 8(4) — ICT Asset Identification. The static capability inventory **SHALL** include: the document extraction model (provider, version, supported document types), the vision/OCR model, the assessment agent model, the policy terms database, the fraud scoring service (provider, version, API specifications), the actuarial engine (internal or third-party, version, calculation methodology), and the payment initiation system. The dynamic runtime inventory **SHALL** log, per claim: which document types were processed, which data sources were queried, whether the actuarial engine was invoked (and in what mode—read-only coverage verification versus risk recalculation), and the disposition pathway followed.

Article 10 — Detection. Three-tier monitoring **SHALL** track: *Envelope compliance*: settlement amounts within approved limits, claim types within approved scope, actuarial engine access mode (read-only versus risk recalculation—Boundary B3 monitoring). *Behavioural distribution*: approval rate by claim category and policyholder demographic, average settlement amount distribution, denial rate, fraud referral rate, processing time, and actuarial engine invocation frequency. *Outcome quality*: policyholder complaint rate, appeal success rate (proportion of denied claims overturned on human review), settlement accuracy (comparison of agent settlements to retrospective human adjuster assessments on sampled claims), and demographic parity metrics across settlement outcomes.

Article 19 — Incident Classification. The following agent-specific events **SHALL** be assessed: systematic settlement bias detected through demographic parity monitoring (cumulative onset), actuarial engine invocation in risk recalculation mode without high-risk system compliance infrastructure active (ART-3 boundary crossing—onset at first unauthorised recalculation), OCR extraction error rate exceeding baseline (CRT-1 cascade precursor), and payment initiation triggering downstream account freezes (ART-5 state propagation).

5.4.6 Safe Halt Specification

1. **Immediate capability suspension:** both document extraction and assessment agents are prevented from processing new claims.
2. **In-progress claim preservation:** claims currently being processed are saved with their current state (documents extracted, data gathered, preliminary assessment, pending disposition).
3. **Payment hold:** any settlement payments authorised but not yet executed are frozen. Payments already submitted to the payment system are flagged for monitoring but cannot be recalled if already processed.

4. **Policyholder communication hold:** any pending communications to policyholders (approval letters, denial notices, information requests) are frozen for human review before dispatch.
5. **Regulatory deadline assessment:** claims approaching statutory response deadlines (many jurisdictions impose maximum response times for insurance claims) are flagged for immediate human attention.
6. **Halt state log:** complete state captured including all in-progress claims, their processing stage, pending payments, pending communications, and regulatory deadline status.

5.5 Profile D: Customer Onboarding Agent

5.5.1 Architecture Reference

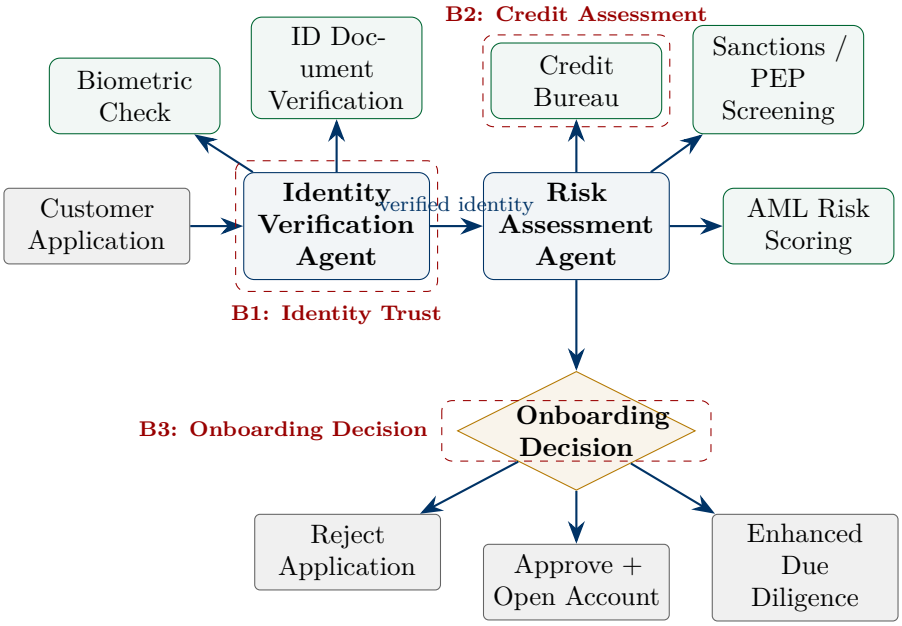

Figure 7: Customer onboarding agent. Three compliance-critical boundaries: B1 (identity trust—the downstream risk assessment depends entirely on the identity verification agent's output), B2 (credit assessment—querying a credit bureau for onboarding risk scoring may constitute creditworthiness assessment under Annex III point 5(b)), B3 (onboarding decision—autonomous approval, rejection, or enhanced due diligence determination with legal effects on the applicant).

5.5.2 Autonomy Tier and Oversight Model

Autonomy classification: Tier 2 for standard onboarding within defined acceptance criteria. Tier 3 if the agent autonomously determines acceptance criteria, adjusts risk thresholds based on portfolio composition, or handles borderline applications without escalation.

Intervention models:

- **Model 1 (Pre-Commitment Gates):** applied to application rejections (all rejections reviewed by human before communication to applicant), enhanced due diligence triggers, and applications involving politically exposed persons (PEPs) or adverse sanctions screening results.
- **Model 2 (Statistical Process Control):** applied to approval rate, rejection rate, enhanced due diligence referral rate, average processing time, and all metrics disaggregated by applicant demographics (nationality, age, gender, postcode) to detect discriminatory patterns.
- **Model 3 (Post-Action Audit):** applied to approved standard-risk applications. Random sampling audit of identity verification accuracy, risk assessment appropriateness, and AML scoring correctness.

5.5.3 Applicable Risk Categories

Risk Category	Severity	Customer Onboarding Manifestation
ART-3 Fiduciary Boundary	Critical	Risk assessment agent queries a credit bureau to assess onboarding risk. The query and its use in the acceptance decision may constitute creditworthiness assessment under Annex III point 5(b), reclassifying the system from non-high-risk onboarding to high-risk credit scoring. The boundary is interpretive: using credit data to assess "onboarding risk" versus using credit data to assess "creditworthiness" may be functionally identical but regulatorily distinct.
ART-2 Tool Chain Escalation	Critical	Agent combines identity verification, credit bureau data, sanctions screening, and AML risk scoring into a comprehensive applicant profile. The assembled profile constitutes a detailed risk dossier exceeding the scope of any individual tool's assessment—a capability the institution did not explicitly authorise.
ART-1 Decision Drift	High	Agent's acceptance threshold shifts through accumulated applicant exposure. In periods of high application volume with predominantly low-risk applicants, the agent's implicit threshold for "acceptable risk" recalibrates downward, approving applications that earlier iterations would have referred for enhanced due diligence.
CRT-4 Routing Discrimination	High	The risk assessment agent routes applicants to different processing pathways (approve, reject, enhanced due diligence) based on risk scoring. If the routing correlates with protected characteristics —nationality, ethnicity, postcode as proxy for race—the architecture produces structural discrimination.
ART-6 Accountability Void	High	Identity verification agent feeds risk assessment agent which queries credit bureau, sanctions, and AML scoring. When an applicant is rejected, explaining the causal chain—which identity element, which risk factor, which tool output drove the rejection—requires cross-agent reconstruction under GDPR Article 22 and AI Act Article 86 timelines.
ART-5 State Propagation	Medium	Agent opens an account; the account opening triggers KYC periodic review scheduling, product eligibility assessments, and marketing system enrolment—downstream effects the onboarding agent did not directly initiate.

5.5.4 Regulatory Classification

AI Act: The customer onboarding agent operates at the most contested Annex III boundary in financial services. Two analysis paths yield different classifications:

Path 1 — KYC/AML classification: if the onboarding decision is framed as "identity verification and AML risk assessment for account opening," the system is performing a regulatory compliance function (CDD under the AML Directive). This function is not listed in Annex III, and AML-adjacent functions benefit from the financial fraud detection exemption analogy.

Path 2 — Creditworthiness classification: if the onboarding decision incorporates credit bureau data and the acceptance criteria include financial standing, the system is "evaluating the creditworthiness of natural persons" within the meaning of Annex III point 5(b). This triggers high-risk classification with full Chapter III obligations.

The classification turns on the **intended purpose** of the credit bureau query. If the query serves AML risk assessment ("is this applicant's stated income consistent with their transaction

profile?"), it is a CDD function. If the query serves acceptance decisioning based on financial capacity ("can this applicant maintain the account without default risk?"), it is creditworthiness assessment. An autonomous agent that dynamically determines how to use credit bureau data may cross this boundary without the institution's awareness—the paradigmatic ART-3 scenario.

AML Directive: Customer due diligence obligations (Articles 13–14 of Directive (EU) 2015/849) apply to the entire onboarding process. The institution bears responsibility for adequate CDD regardless of whether the process is human-executed or agent-executed. Simplified due diligence (Article 15) and enhanced due diligence (Article 18) decisions must meet the regulatory standard —a standard the institution must demonstrate was met even when the decision was made autonomously.

GDPR Article 22: Directly and acutely engaged. An autonomous rejection of an account application is a decision "based solely on automated processing" that "produces legal effects concerning" the applicant or "similarly significantly affects" them. Article 22(1) prohibits this unless the institution relies on one of the exceptions in Article 22(2): explicit consent, necessity for a contract, or authorisation by EU/Member State law with suitable safeguards. In all cases, the institution must provide meaningful information about the logic involved, the right to human intervention, and the right to contest the decision (Article 22(3)).

DORA: Fully applicable. Customer onboarding supports a critical business function (regulatory compliance and customer acquisition). All third-party tools (ID verification, credit bureau, sanctions screening, AML scoring) are ICT third-party service providers requiring Article 28–30 compliance.

5.5.5 DORA Compliance Requirements

Article 8(4) — ICT Asset Identification. The static capability inventory **SHALL** include: the identity verification agent model, the risk assessment agent model, the ID document verification service (provider, supported document types, jurisdictions), the biometric check service, the credit bureau connection (provider, data fields accessed, query types), the sanctions/PEP screening service, and the AML risk scoring model or service. The dynamic runtime inventory **SHALL** log, per application: which verification checks were performed, which data sources were queried, the query type used for each credit bureau access (CDD versus creditworthiness—Boundary B2 monitoring), and the disposition pathway followed.

Article 10 — Detection. Three-tier monitoring **SHALL** track: *Envelope compliance*: credit bureau query type (CDD-only versus creditworthiness—immediate alert if creditworthiness queries are executed without high-risk compliance infrastructure active), PEP/sanctions hit handling (all hits escalated, no autonomous processing). *Behavioural distribution*: approval rate, rejection rate, and enhanced due diligence rate—all disaggregated by applicant nationality, age, gender, and postcode. Statistical tests for demographic parity (equalised odds, calibration) applied at defined intervals. *Outcome quality*: identity verification false accept rate (accounts opened with fraudulent identity subsequently detected), rejection appeal success rate, and account-level adverse events within 12 months of onboarding (default, fraud, regulatory finding) as a lagging indicator of onboarding quality.

Article 19 — Incident Classification. The following agent-specific events **SHALL** be assessed: systematic rejection bias detected through demographic disaggregation (cumulative onset), credit bureau queries executed in creditworthiness mode without high-risk compliance (ART-3 boundary crossing—onset at first unauthorised query), identity verification false acceptance confirmed through subsequent fraud detection (action-as-incident— onset at the verification decision, detected retrospectively), and sanctions screening hit processed autonomously without escalation (envelope violation—immediate classification).

5.5.6 Safe Halt Specification

1. **Immediate capability suspension:** both identity verification and risk assessment agents are prevented from processing new applications.
2. **In-progress application preservation:** applications currently being processed are saved with their current state (verification steps completed, data gathered, preliminary risk assessment).
3. **Decision hold:** any pending onboarding decisions (approve, reject, enhanced due diligence) are frozen for human review.
4. **Communication hold:** any pending communications to applicants (approval confirmation, rejection notification, information requests) are frozen.
5. **Account activation hold:** accounts approved but not yet activated are held pending human verification.
6. **Regulatory deadline assessment:** applications approaching CDD completion deadlines or customer communication deadlines are flagged for immediate human attention.
7. **Halt state log:** complete state captured including all in-progress applications, verification results, pending decisions, and regulatory deadline status.

The onboarding-specific safe halt concern is **applicant experience continuity**. Unlike internal processing systems, the onboarding agent interacts with prospective customers who expect a response. Prolonged halt without communication damages the institution's reputation and may violate consumer protection requirements for timely response to account applications. The halt state log **SHALL** include applicant communication status, and the institution's halt response procedure **SHALL** include template communications informing applicants of processing delays.

CRSA-1 EU Compliance Series — Forthcoming

Detailed demographic parity monitoring specifications, credit bureau query classification methodology for distinguishing CDD from creditworthiness assessment, and GDPR Article 22 compliance templates for autonomous onboarding decisions will be provided in implementation guidance accompanying the CRSA-1 Financial Services Edition.

6 DORA Register of Information for Agent Tool Connections

6.1 The Registration Problem

Article 28(3) of DORA requires financial entities to maintain a Register of Information of *all* contractual arrangements on the use of ICT services provided by ICT third-party service providers. The Implementing Technical Standards (Commission Implementing Regulation (EU) 2024/2956) specify a multi-table relational data model with over 60 mandatory fields across six tables. The first submission deadline was 30 April 2025. There is no de minimis threshold: every ICT service arrangement must be registered.

For non-agentic AI systems, registration is straightforward. A RAG pipeline using an OpenAI API and a vector database has two third-party service arrangements: one with OpenAI (or Microsoft, if accessed through Azure OpenAI Service) and one with the vector database provider. Each arrangement receives entries in tables B.02 (contractual arrangement), B.03 (provider identification), and B.04 (service description and criticality).

For agentic systems, the registration problem is qualitatively different. An agent with access to 15 MCP tools, three external data sources, two communication platforms, and a payment system has potentially 21 ICT third-party service arrangements. The tool inventory may change as the institution connects or disconnects tools. The agent may invoke different subsets of tools for different tasks, making the "criticality" of each tool context-dependent rather than fixed.

And the agent may autonomously prefer certain tools over others, creating runtime dependencies that differ from the designed dependencies.

No guidance from the ESAs, the Commission, or any national supervisor addresses how agentic tool connections should be mapped to the Register of Information ITS template. This section provides the Tool Inventory Registration Methodology that resolves the ambiguity.

6.2 ITS Template Structure and Agent Mapping

The Register of Information ITS specifies six tables. Four are directly relevant to agent tool registration:

Table B.02 — Contractual Arrangements. Each contractual arrangement with an ICT service provider requires an entry recording the arrangement type, start date, renewal terms, governing law, and the business functions supported. For agents, the key question is *granularity*: does each tool connection constitute a separate contractual arrangement, or can multiple tools be consolidated?

Table B.03 — Provider Identification. Each ICT third-party service provider requires an entry recording the provider's LEI (where available), jurisdiction, corporate structure, and relationship to the financial entity. For agents accessing tools from the same provider through different interfaces (e.g., multiple endpoints from the same SaaS vendor), a single B.03 entry with multiple linked B.02 entries is appropriate.

Table B.04 — Service Description and Criticality. Each ICT service requires a description and a criticality assessment (critical or important function, or not). For agents, tool criticality depends on the business function the tool supports *through the agent*—the same tool may be critical when used by one agent (e.g., a credit bureau queried by a high-risk credit scoring agent) and non-critical when used by another (e.g., the same credit bureau queried by an internal research tool).

Table B.06 — Sub-Outsourcing. For critical or important functions, the register must identify sub-outsourcing chains. When an agent accesses a tool through an intermediary (e.g., an MCP server hosted by a platform provider that in turn calls an external API), the sub-outsourcing chain must be documented.

6.3 Tool Inventory Registration Methodology

This specification defines three registration approaches, each appropriate to different institutional contexts. Financial entities **SHALL** select the approach that provides the most accurate representation of their agent's tool dependencies while maintaining operational feasibility.

6.3.1 Approach 1: Provider-Level Registration

Provider-Level Registration

All tools from the same provider are consolidated under a single contractual arrangement entry (B.02), linked to one provider entry (B.03). The service description (B.04) covers the full set of tools accessible from that provider, with criticality assessed at the highest criticality level of any business function supported by any tool from that provider.

Advantages: lowest operational burden; aligns with how contracts are typically structured (one agreement per vendor); reduces the number of B.02 entries.

Limitations: may overstate criticality (if one tool from a provider supports a critical function, all tools from that provider inherit the criticality classification and its Article 30(3) contractual requirements); obscures tool-level dependency granularity; does not capture the agent's differential reliance on specific tools from the same provider.

Appropriate when: the institution has relatively few tool providers (fewer than 10), each provider supplies a coherent set of related tools, and the institution accepts conservative criticality classification.

6.3.2 Approach 2: Function-Level Registration

Function-Level Registration

Tools are grouped by the business function they support through the agent. Each business function generates one contractual arrangement entry (B.02) covering all tools that the agent uses to fulfil that function, regardless of how many providers supply those tools. Criticality is assessed at the function level, directly aligned with DORA's function-based criticality model.

Advantages: aligns directly with DORA's function-based architecture; criticality assessment is natural (the function is either critical or not); captures the agent's operational purpose rather than its technical topology.

Limitations: a single tool supporting multiple business functions appears in multiple B.02 entries, creating redundancy; requires restructuring when the agent's functional allocation changes; may create multiple B.02 entries linked to the same B.03 provider entry, complicating the register's relational structure.

Appropriate when: the institution's agents support clearly defined business functions with stable tool allocations, and the institution prioritises regulatory alignment over operational simplicity.

6.3.3 Approach 3: Capability-Level Registration

Capability-Level Registration

Each tool in the agent's capability inventory receives its own contractual arrangement entry (B.02), with independent criticality assessment based on the most critical business function that tool supports. This is the most granular approach and provides the highest fidelity representation of the agent's dependency landscape.

Advantages: highest dependency visibility; tool-level criticality enables precise Article 30(3) contract negotiation (only genuinely critical tools require full contractual provisions); supports tool-level exit strategy planning; enables precise concentration risk assessment at tool granularity.

Limitations: highest operational burden; an agent with 20 tools requires 20 B.02 entries with associated B.03 and B.04 entries; updates to the tool inventory require register updates before the agent accesses the new tool.

Appropriate when: the institution deploys Tier 3 agents with large tool inventories, regulatory scrutiny is anticipated, and the institution requires maximum transparency for supervisory review.

6.4 Registration Methodology Selection

Factor	Provider-Level	Function-Level	Capability-Level	Recommendation
Tool inventory size <10	Adequate	Adequate	Optimal	Any approach feasible
Tool inventory size 10–30	Adequate	Optimal	Burdensome	Function-level preferred
Tool inventory size >30	Adequate	Optimal	Impractical	Function-level with provider consolidation
Tier 2 agents	Adequate	Adequate	Optional	Provider-level sufficient
Tier 3 agents	Insufficient	Adequate	Optimal	Capability-level or function-level
Multi-agent systems	Insufficient	Adequate	Optimal	Function-level minimum; capability-level for supervisory transparency

Table 4: Registration methodology selection matrix. "Adequate" indicates the approach meets regulatory requirements. "Optimal" indicates the approach provides the best balance of compliance and operational value. "Insufficient" indicates the approach does not provide adequate dependency visibility for the deployment context.

The general principle is that registration granularity **SHOULD** increase with agent autonomy. Provider-level registration is sufficient for Tier 2 agents with bounded tool inventories. Function-level registration is the recommended default for most deployments. Capability-level registration is appropriate for Tier 3 agents and for institutions expecting supervisory scrutiny of their agent deployments.

6.5 Dynamic Tool Inventory Management

The Register of Information is not a static filing. Article 28(3) requires it to be "maintain[ed] and update[d]." For agentic systems, three dynamic inventory management requirements arise:

Pre-access registration. When a new tool is added to the agent's capability inventory, the register **SHALL** be updated *before* the agent's first access to that tool. This prevents the agent from creating unregistered ICT dependencies. The tool governance process **SHALL** include register update as a mandatory step in the tool activation workflow.

Deactivation recording. When a tool is removed from the agent's capability inventory, the register entry **SHALL** be updated to reflect the deactivation date and reason. Historical entries are retained for audit trail purposes per Article 28(3)'s maintenance requirement.

Runtime divergence monitoring. The agent's actual tool usage **SHALL** be continuously compared against the registered capability inventory. Any access to a tool not in the registered inventory constitutes an anomaly requiring investigation under Article 10 (detection). Any registered tool that the agent has not accessed within a defined period **SHOULD** be reviewed for continued necessity—unused tools in the capability inventory represent latent dependencies that inflate concentration risk assessments without providing operational value.

6.6 Concentration Risk Assessment for Agent Tool Portfolios

Article 28(4)(c) requires assessment of whether contractual arrangements "may contribute to reinforcing ICT concentration risk." For agents, concentration risk must be assessed across three dimensions:

Provider concentration. Multiple tools from the same provider, or from providers with common ownership or infrastructure dependencies. Standard DORA concentration analysis applies, extended to the agent's full tool inventory. OpenAI and Azure OpenAI Service **SHOULD** be treated as "closely connected" providers under Article 29(1)(b).

Capability concentration. The agent's critical functions depend on tools for which no substitute exists in the capability inventory. If the agent's AML investigation function requires a specific sanctions screening service with no registered alternative, this is a capability concentration that the exit strategy must address.

Autonomous concentration. The agent autonomously prefers certain tools over registered alternatives, creating runtime concentration that the institution did not design. This is ART-4 (Autonomous Concentration) applied at the tool level. Detection requires monitoring the agent's tool utilisation distribution and alerting when usage concentrates on a single provider beyond defined thresholds.

The concentration risk assessment **SHALL** be updated at minimum annually and upon any material change to the agent's tool inventory. For Tier 3 agents, the assessment **SHOULD** include autonomous concentration analysis based on runtime tool utilisation data from the preceding period.

6.7 Sub-Outsourcing Chains for MCP-Based Agents

When an agent accesses tools through MCP servers, the sub-outsourcing chain (Table B.06) may have multiple layers:

1. **Layer 1 — MCP server provider.** The entity operating the MCP server through which the agent connects. This may be the institution itself (for internally hosted MCP servers), a platform provider (for hosted MCP infrastructure), or the tool provider directly (if the tool provider operates its own MCP endpoint).

2. **Layer 2 — Underlying service provider.** The entity providing the actual service accessed through the MCP server. If the MCP server proxies requests to an external API, the external API provider is a sub-outsourcing party.

3. **Layer 3 — Infrastructure provider.** The cloud platform hosting the MCP server and/or the underlying service. This layer may introduce additional concentration risk if multiple MCP servers are hosted on the same cloud platform.

For critical or important functions, Table B.06 requires documentation of the sub-outsourcing chain through all layers. The institution **SHALL** obtain sufficient information from the MCP server provider to identify Layer 2 and Layer 3 dependencies. Where the MCP server provider cannot or will not disclose the sub-outsourcing chain, this information gap **SHALL** be documented in the concentration risk assessment and compensating measures applied (such as treating the tool as a single point of failure in the exit strategy).

CRSA-1 EU Compliance Series — Forthcoming

Worked examples mapping specific agentic tool inventories to the ITS template across all three registration approaches, including sample B.02, B.03, B.04, and B.06 entries for common financial services agent configurations, will be provided in implementation guidance accompanying the CRSA-1 Financial Services Edition.

7 Incident Classification for Autonomous Agent Actions

7.1 The Onset Problem

DORA's incident classification framework (Article 18; CDR 2024/1772) assumes incidents with identifiable onset: a system fails, an anomaly is detected, an intrusion is discovered. The classification criteria evaluate the incident's *impact* against six materiality thresholds, and the reporting timeline runs from *classification* (4 hours for initial notification, 72 hours for intermediate report, 1 month for final report). The framework operates on the implicit assumption that the financial entity can identify *when* an incident began.

For autonomous agent actions, this assumption fails in three distinct ways, identified in Section 2: action-as-incident (a single autonomous action that causes harm), cumulative onset (a pattern of individually reasonable actions that collectively cause harm), and delegation chain attribution (harm arising from a multi-agent interaction where no single agent's action is independently harmful). This section provides the Agent Incident Classification Decision Logic that resolves the onset ambiguity for each scenario.

7.2 Agent Incident Classification Decision Logic

The decision logic operates in two stages. Stage 1 determines *whether* an agent-related event constitutes an ICT-related incident. Stage 2 applies the standard CDR 2024/1772 materiality criteria to determine *whether* the incident is major.

7.2.1 Stage 1: Incident Determination

The following event categories **SHALL** be assessed as potential ICT-related incidents for agentic systems:

Category A — Discrete Agent Action. A single identifiable agent action that produces an adverse outcome. Examples: the agent executes a trade at a price materially worse than the

benchmark, denies a valid insurance claim, approves a fraudulent account application, or closes an AML alert that is subsequently confirmed as genuine suspicious activity.

Onset determination: the onset is the timestamp of the agent's action. This is the simplest case and maps directly to DORA's standard incident model. The action is logged with a timestamp; the incident management process evaluates the action against defined adverse outcome criteria.

Detection mechanism: real-time outcome monitoring (Intervention Model 2) or post-action audit (Intervention Model 3) identifies the adverse action. The detection latency—the time between the action and its identification as adverse—determines how quickly the incident management process can be initiated.

Category B — Cumulative Behavioural Pattern. A pattern of agent actions that are individually within normal parameters but collectively produce a harmful outcome. Examples: systematic under-settlement of claims from a specific demographic (ART-1 Decision Drift), gradual increase in alert closure rate without corresponding decrease in false positive rate (ART-1), or progressive concentration of trade execution on a single venue (ART-4 Autonomous Concentration).

Onset determination: the onset is the point at which the cumulative pattern *became statistically detectable*—the earliest point at which the institution's monitoring system could have identified the pattern given its configured detection thresholds and monitoring frequency. This is necessarily retrospective: the pattern is detected at time T_d but the onset is determined to be time $T_o < T_d$ through statistical analysis.

Detection mechanism: behavioural distribution monitoring (Intervention Model 2) detects the pattern through control chart breach or statistical test failure. The **attribution window**—the maximum lookback period the institution applies to determine onset—**SHALL** be defined in the incident management policy and **SHALL** not exceed 90 days. Beyond 90 days, the institution's monitoring system should have detected the pattern; failure to detect within 90 days is itself a monitoring deficiency requiring separate assessment.

Reporting timeline: the 4-hour initial notification clock starts at the point of *classification*, not at the retrospectively determined onset. The institution classifies at T_d; the notification is due within 4 hours of classification. The intermediate report (72 hours) and final report (1 month) **SHALL** include the retrospectively determined onset T_o and the explanation of why detection occurred at T_d rather than earlier.

Category C — Delegation Chain Failure. An adverse outcome arising from the interaction between multiple agents or between an agent and its tools, where no single agent's action is independently harmful but the chain produces harm. Examples: a triage agent deprioritises an alert that the investigation agent would have escalated had it been prioritised earlier; a document extraction agent misreads a field that the assessment agent processes correctly given its input, producing an incorrect settlement.

Onset determination: the onset is the **earliest contributing action** in the delegation chain—the first agent action that, if performed correctly, would have prevented the adverse outcome. This requires causal reconstruction through the distributed log infrastructure described in ART-6 mitigation.

Detection mechanism: outcome monitoring detects the harm; causal reconstruction (using the distributed correlation framework from Section 4, ART-6) traces the causal chain to identify contributing actions. Detection latency includes both the time to detect the harm and the time to reconstruct the causal chain.

Reporting timeline: the 4-hour clock starts at classification. The intermediate report **SHALL** include the causal chain reconstruction identifying each contributing agent and action. If causal

reconstruction is not completable within 72 hours, the intermediate report **SHALL** state the reconstruction status and provide the complete chain in the final report.

7.2.2 Stage 2: Materiality Assessment

Once an event is determined to be an ICT-related incident under Stage 1, the standard CDR 2024/1772 materiality criteria apply. Two of six criteria must be met for major classification. The following agent-specific interpretive guidance applies:

Clients affected. For Category B incidents (cumulative patterns), the client count is the total number of clients affected by the pattern across the full attribution window, not merely those affected at the point of detection. A systematic settlement bias affecting 200 clients per week over a 12-week attribution window affects 2,400 clients—potentially meeting the >10% threshold even if no single week's affected population does.

Duration. For Category B incidents, duration runs from the retrospectively determined onset T_o to the point of remediation. The duration may be weeks or months, substantially exceeding the 24-hour threshold. For Category A incidents, duration is measured from the agent's action to the completion of remediation (including reversal of harmful actions where feasible).

Economic impact. For agent-related incidents, the economic impact includes: direct financial loss (incorrect settlements, erroneous trades), remediation cost (reversal transactions, customer compensation), and regulatory exposure (potential fines, supervisory action). The economic impact of Category B incidents accumulates across the attribution window and may substantially exceed the impact visible at any single point.

Malicious unauthorised access. Agent-related incidents involving adversarial manipulation (prompt injection, tool output poisoning, goal manipulation) **SHALL** be assessed as potential "malicious unauthorised access" triggering automatic major classification. The attacker uses the agent as a vector; the agent's autonomous actions constitute the "access" that the attacker directs. This interpretation aligns with DORA's technology-neutral framework: the attack surface includes the agent's decision-making process, not merely its network interfaces.

7.3 Dual-Reporting Under AI Act Article 73(9)

For agents classified as high-risk under the AI Act (credit scoring agents under Annex III point 5(b), life/health insurance pricing agents under point 5(c)), incident reporting obligations arise under both DORA and the AI Act. Article 73(9) resolves the overlap: for financial entities subject to DORA, AI Act serious incident reporting is limited to incidents involving **fundamental rights infringements** (Article 3(49)(c)). All other incident categories (system failure, health/safety harm, infrastructure disruption) are reported exclusively under DORA.

For agent-related incidents, the fundamental rights trigger is most likely to be engaged by:

1. **Discriminatory patterns** (Category B): systematic bias in the agent's decisions affecting protected groups constitutes a fundamental rights infringement requiring AI Act reporting in addition to DORA reporting.
2. **Automated decision-making without safeguards** (Category A): an agent autonomously making a decision with legal effects on a natural person without the GDPR Article 22 safeguards being active constitutes a fundamental rights infringement.
3. **Fiduciary boundary violation** (ART-3): an agent that autonomously crosses from a non-high-risk function into a high-risk function and makes decisions in the high-risk domain without the required oversight constitutes operation of an unregistered high-risk AI system—a regulatory violation that may implicate fundamental rights.

For these scenarios, the institution must report under both DORA (4-hour initial notification to the financial supervisor) and the AI Act (to the market surveillance authority, which under

Article 74(6) is the financial supervisor itself). In practice, this means a single supervisor receives both reports, but the reports serve different regulatory purposes and must contain the information required by each framework.

7.4 Mandatory Review Triggers

Independent of the incident classification process, the following conditions **SHALL** trigger mandatory review of the agent's operations, regardless of whether an incident has been classified:

1. **Action envelope breach**: any agent action outside the approved action envelope triggers immediate review. A single breach may not constitute an incident, but a pattern of breaches constitutes a Category B event.
2. **Escalation rate anomaly**: the agent's escalation rate deviates beyond defined control limits (either direction— too few escalations may indicate escalation suppression, too many may indicate model degradation).
3. **Demographic parity alert**: statistical tests for demographic parity in the agent's decisions fail at the defined significance level.
4. **Tool inventory divergence**: the agent accesses a tool not in the registered capability inventory, or a registered tool that has not been accessed within the defined activity window shows sudden usage.
5. **Autonomous concentration alert**: the agent's tool utilisation distribution concentrates beyond defined thresholds.
6. **Causal reconstruction failure**: a routine causal reconstruction test (ART-6 detection methodology) fails to reconstruct the complete decision chain within the required timeframe.

8 Agent Exit Strategy Framework

8.1 The Agent Switching Cost Problem

DORA Article 28(8) requires financial entities to develop and maintain exit strategies for ICT third-party dependencies, including "transition plans" and provisions ensuring "that the financial entity can withdraw from those contractual arrangements without disruption to its business activities." For traditional ICT services, exit strategies are demanding but architecturally straightforward: migrate data, switch providers, validate functionality.

For agentic AI systems, exit strategies face an exponential switching cost problem. An agent's behaviour is not merely a function of its model—it is an emergent property of the model, the prompt architecture, the tool integration logic, the accumulated context, the action envelope definition, the monitoring baselines, and the institutional knowledge embedded in the system's configuration. Each of these elements is, to varying degrees, model-specific.

The switching costs for agent components include:

1. **Model substitution cost.** Replacing the foundation model requires re-engineering prompts, re-running all evaluation benchmarks, and re-establishing accuracy baselines. Semantic equivalence between models cannot be guaranteed—a prompt that produces correct behaviour with GPT-4 may produce incorrect behaviour with Claude, not because either model is defective but because they interpret instructions differently.

2. **Prompt architecture cost.** The agent's prompt templates, few-shot examples, chain-of-thought scaffolding, and system instructions constitute institutional intellectual property developed through iterative refinement. These are model-specific: optimal prompting strategies differ between model families. Migration requires rebuilding the prompt architecture, which may take weeks to months of engineering effort.

3. **Tool integration cost.** The agent's parsing of tool outputs, error handling for tool failures, and response interpretation logic may be tuned to specific tool provider output formats. Switching a tool provider requires updating the integration logic and re-validating the agent's behaviour with the new tool's output characteristics.

4. **Evaluation re-establishment cost.** The agent's performance baselines, control chart parameters, action envelope boundaries, and monitoring thresholds were calibrated for the current configuration. After any component substitution, all baselines must be re-established through a calibration period, during which the agent may operate with reduced monitoring confidence.

5. **Regulatory re-validation cost.** For high-risk agents, any material change to the agent's configuration may constitute a substantial modification under AI Act Article 3(23), requiring conformity re-assessment. The exit itself may trigger the compliance obligations it was designed to maintain continuity through.

6. **Accumulated context cost.** Agents that maintain persistent memory across invocations accumulate operational context that is model-specific in format and interpretation. Migrating accumulated context to a new model requires semantic translation with no guarantee of fidelity.

The compound effect is that exiting from a foundation model provider for an agentic system is not a configuration change—it is a partial system rebuild with regulatory re-validation. DORA's exit strategy requirement must be interpreted with this reality in mind.

8.2 The Portable Agent Specification

This specification proposes the **Portable Agent Specification (PAS)** as the architectural pattern that enables credible exit strategies for agentic systems. The PAS captures the agent's *behavioural contract*—the specification of what the agent does, how it is governed, and how its performance is measured—in a model-independent format.

The PAS does not make model substitution costless. It makes model substitution *structurally feasible* by ensuring that the institution retains all information necessary to rebuild the agent on an alternative foundation, and by defining the validation criteria the rebuilt agent must satisfy before operational deployment.

The PAS **SHALL** contain the following elements:

1. **Behavioural specification.** A model-independent description of the agent's intended behaviour: the tasks it performs, the decisions it makes, the quality criteria for correct execution, and the boundary conditions defining acceptable versus unacceptable performance. Written in terms of inputs, expected outputs, and acceptance criteria—not in terms of model-specific implementation.

2. **Tool interface specification.** The abstract interface each tool must provide (input format, output format, error responses, latency requirements), independent of the current provider's specific API. This enables tool-level exit: any provider implementing the abstract interface can substitute for the current provider.

3. **Action envelope definition.** The complete specification of the agent's approved action space, escalation boundaries, and pre-commitment gates. These governance parameters are model-independent and transfer directly to any rebuilt agent.

4. **Evaluation benchmark suite.** The test cases, evaluation datasets, and acceptance criteria used to validate the agent's performance. The benchmark suite is the primary instrument for verifying that a rebuilt agent on an alternative foundation satisfies the behavioural specification. The suite **SHALL** include: functional correctness tests (does the agent complete tasks correctly?), envelope compliance tests (does the agent respect the action envelope?), robustness tests (does the agent handle tool failures and adversarial inputs?), and fairness tests (does the agent produce equitable outcomes across demographic groups?).

5. **Monitoring specification.** The KRIs, control chart parameters, alert thresholds, and monitoring infrastructure requirements, specified in terms of the metrics to be tracked rather than the implementation mechanism. The monitoring specification transfers to the rebuilt agent, enabling rapid re-establishment of the three-tier monitoring framework.

6. **Regulatory compliance profile.** The agent's AI Act classification, applicable DORA provisions, sectoral regulatory obligations, and conformity assessment status. The compliance profile identifies which regulatory validations must be re-executed upon model substitution and which transfer unchanged (action envelope governance, logging requirements, oversight architecture).

8.3 Exit Strategy Components

A credible exit strategy for an agentic system **SHALL** address each of the following components:

Alternative provider mapping. For each component in the agent's architecture (foundation model, each tool, each data source), identify at minimum one alternative provider. The mapping **SHALL** include: the alternative provider's capability profile, known compatibility issues with the PAS tool interface specifications, estimated migration effort, and any contractual or licensing constraints.

Migration procedure. A documented, tested procedure for rebuilding the agent on the alternative foundation. The procedure **SHALL** include: prompt architecture migration methodology (systematic re-engineering rather than copy-paste), tool integration re-validation steps, evaluation benchmark execution against the PAS acceptance criteria, monitoring baseline re-establishment protocol, and regulatory re-assessment trigger analysis (determining whether the migration constitutes a substantial modification).

Transition period specification. The estimated duration of the migration, including parallel running (both the current and alternative agents operating simultaneously with output comparison), performance validation, and cutover. The transition period **SHALL** be specified in the Article 30(3)(f) contractual exit provisions.

Fallback mode. The operational mode during transition if the current provider becomes unavailable before migration is complete. Options include: degradation to Tier 1 (assistive mode with human execution), fallback to rule-based processing for critical functions, or activation of an alternative agent maintained in warm standby on a different foundation model.

Data portability. Specification of which data elements the institution must be able to extract from the current provider: fine-tuning datasets, evaluation results, accumulated context logs, and performance baselines. Article 30(2)(a) requires contractual provisions for data return; the exit strategy **SHALL** specify what "data" means for an agentic system.

8.4 Exit Testing

DORA does not explicitly require testing of exit strategies, but BaFin's December 2025 guidance requires business continuity testing that implicitly encompasses exit scenario validation. For agentic systems, exit testing **SHALL** include:

1. **PAS completeness verification.** Confirm that the Portable Agent Specification contains all information necessary to rebuild the agent without reference to the current implementation. A third party (internal team unfamiliar with the current agent, or external assessor) should be able to understand the agent's intended behaviour, governance, and validation criteria from the PAS alone.

2. **Alternative provider validation.** Execute the PAS evaluation benchmark suite against the alternative provider's model (or a representative sample) to verify that acceptable performance is achievable. This need not be a full migration—a proof-of-concept demonstrating that the behavioural specification can be satisfied on the alternative foundation is sufficient for exit readiness.

3. **Fallback mode activation.** Test the transition to fallback mode under simulated provider unavailability. Verify that critical business functions continue to be served during fallback and that the transition does not create state corruption (ART-5) or accountability voids (ART-6).

4. **Timeline validation.** Verify that the estimated migration duration is realistic by executing the first stages of the migration procedure (prompt re-engineering, tool integration for one or two tools) and extrapolating.

Exit testing **SHOULD** be conducted annually for Tier 3 agents supporting critical business functions and upon any material change to the agent's architecture or tool inventory.

Honest Framing

The Portable Agent Specification reduces exit risk but does not eliminate switching costs. Model substitution for an agentic system will always require engineering effort, re-validation, and a transition period. The PAS ensures that the institution retains the *knowledge* necessary for substitution and the *criteria* for validating the substituted system. It does not guarantee that the substituted agent will perform identically—model capabilities are not fungible, and behavioural equivalence across model families cannot be assured. The exit strategy accepts this residual risk and manages it through parallel running, benchmark validation, and fallback modes.

9 What This Framework Cannot Guarantee

Honest Framing

This specification provides governance infrastructure for autonomous AI agents deployed in regulated financial services. It provides risk reduction, accountability mechanisms, and compliance methodology. It does not provide behavioural safety guarantees. This section states explicitly what the framework cannot do.

9.1 The Irreducible Tension

Autonomous agency and complete human control are mutually exclusive. An agent that requires human approval for every action is not autonomous; it is an assistive tool. An agent that acts

without per-action approval creates a temporal window during which it operates without effective human oversight. The Agent Oversight Architecture (Section 3) manages this tension through tier-appropriate intervention models, but it does not resolve it. The tension is architectural, not procedural—no governance process eliminates it without eliminating the autonomy.

Financial institutions deploying autonomous agents accept this tension. The management body's approval of an agentic system under DORA Article 5 constitutes acceptance of a non-zero probability that the agent will take autonomous actions producing adverse outcomes before detection. The framework's function is to minimise that probability, bound the severity of adverse outcomes, and ensure accountability when they occur. It does not guarantee their prevention.

9.2 What the Framework Provides

1. **Risk identification.** The Agentic Risk Taxonomy (ART-1 through ART-6) provides a vocabulary for identifying risks that no existing framework names. Financial institutions cannot govern risks they cannot identify; the taxonomy makes agent-specific risks visible and assessable.

2. **Regulatory mapping.** The DORA gap analysis (Section 2) and AI Act oversight paradox analysis (Section 3) identify where existing regulation fails for agents and provide interpretive guidance for applying those regulations in the agentic context. This does not create new law; it provides the compliance interpretation necessary for institutions operating in a regulatory vacuum.

3. **Governance architecture.** The three intervention models, the three-tier monitoring framework, the action envelope methodology, and the safe halt specifications provide structural governance mechanisms. These mechanisms reduce the probability and severity of adverse agent actions but do not eliminate them.

4. **Operational methodology.** The Tool Inventory Registration Methodology (Section 6), Agent Incident Classification Decision Logic (Section 7), and Portable Agent Specification (Section 8) provide procedures that compliance teams can implement. These procedures operationalise governance but depend on correct implementation and continuous maintenance.

5. **Accountability infrastructure.** The distributed causal logging framework (ART-6 mitigation), the incident attribution protocols, and the mandatory review triggers ensure that when adverse outcomes occur, the institution can reconstruct what happened, why, and what should change. Accountability does not prevent harm; it ensures that harm produces institutional learning and regulatory transparency.

9.3 What the Framework Cannot Provide

1. **Behavioural safety guarantees.** No governance framework can guarantee that an autonomous agent will not take harmful actions. The agent's behaviour emerges from the interaction of a non-deterministic model, a dynamic environment, and a variable tool landscape. Formal verification of agent behaviour is not achievable for current foundation-model-based agents operating in open environments.

2. **Regulatory certainty.** This specification provides interpretive guidance for applying DORA and the AI Act to agentic systems. It does not constitute regulatory guidance from a competent authority. The ESAs, the Commission, and national supervisors have not published authoritative guidance on agent governance. Interpretations in this specification may diverge from future regulatory positions. The specification identifies where authoritative guidance is needed and who must provide it.

3. **Complete risk enumeration.** The ART taxonomy captures agent-specific risks identifiable as of March 2026. Agentic architectures are evolving rapidly. Novel risk categories will emerge as agents acquire longer planning horizons, broader tool access, and multi-agent coordination capabilities. The taxonomy is designed for extensibility but cannot anticipate every future risk.

4. **Model-specific implementation.** The framework is model-agnostic by design. It does not specify how to implement action envelope enforcement for GPT-4 versus Claude versus Llama, or how to calibrate monitoring thresholds for specific model families. Implementation requires model-specific engineering that this specification deliberately does not provide, to maintain generality and to avoid premature freezing of implementation choices in a rapidly evolving technology landscape.

5. **Elimination of the oversight gap.** For Tier 3 agents, there exists an irreducible temporal window between the agent's action and the oversight system's detection of that action. The framework minimises this window through real-time monitoring and maximises reversal capability through transactional architecture, but the window cannot be reduced to zero without eliminating autonomy. Any institution deploying a Tier 3 agent accepts this window as a residual risk.

9.4 The Proportionality Question

DORA applies proportionally to the size and risk profile of the financial entity (Article 4; simplified framework under Article 16). The AI Act applies proportionally to the risk classification of the AI system, not to the size of the deploying entity. This specification applies in full to any financial entity deploying agentic AI systems in functions subject to DORA and the AI Act.

The proportionality tension is real: a small payment institution deploying a Tier 2 AML investigation agent faces the same ART risk categories as a G-SIB deploying the same architecture. The risks are the same; the institution's capacity to implement the governance framework differs. This specification does not resolve the proportionality tension. It provides the complete framework; the institution must calibrate implementation depth to its risk profile, regulatory expectations, and operational capacity. The three autonomy tiers (Section 1.3) provide the primary calibration mechanism: Tier 1 agents require minimal governance extensions, Tier 2 agents require the core framework, Tier 3 agents require the complete framework. Institutions deploying only Tier 1 or Tier 2 agents may apply the framework selectively.

BaFin's December 2025 guidance explicitly excludes entities operating under DORA's simplified framework (Article 16) from its AI guidance scope. This specification does not exclude any entity. The risks that agentic systems create do not diminish because the deploying institution is small. A small institution that cannot implement the full framework should assess whether it can deploy agentic systems responsibly—not whether the framework can be simplified to accommodate its constraints.

A Agent Capability Declaration Template

This template provides the structured format for documenting an agentic system's capabilities for Annex IV technical documentation (AI Act Article 11) and DORA ICT asset identification (Article 8(4)). Financial entities **SHALL** complete this declaration for each deployed agentic system.

A.1 System Identification

A.1 — Agent Identity

(a) System name and version identifier.
(b) Autonomy tier classification (Tier 1, Tier 2, or Tier 3) with justification referencing the five agentic characteristics in Section 1.2.
(c) Applicable intervention models (Model 1, Model 2, Model 3, or combination) with justification.
(d) Business function(s) supported and DORA criticality classification for each.
(e) AI Act risk classification (high-risk under Annex III, limited-risk, or not classified) with classification rationale.
(f) Date of initial deployment and date of most recent material change.

A.2 Capability Envelope

A.2 — Tool and Data Source Inventory

(a) Complete enumeration of tools, APIs, data sources, and external systems the agent can access. For each: provider identity, service description, access mode (read-only or read-write), data types exchanged, and DORA Register of Information entry reference (B.02 identifier).
(b) Foundation model identification: provider, model name and version, access method (API, on-premise, cloud-hosted), and contractual arrangement reference.
(c) MCP server inventory (if applicable): server provider, hosted tools, sub-outsourcing chain, and B.06 entry reference.

A.3 Action Envelope

A.3 — Approved Autonomous Action Space

(a) Complete specification of actions the agent may execute without per-action human approval, including any parametric boundaries (amount thresholds, scope limitations, instrument restrictions).
(b) Escalation boundary definition: conditions under which the agent must escalate to human oversight, enumerated exhaustively.
(c) Pre-commitment gate specification: irreversible actions requiring human approval, with the approval mechanism described.
(d) Regulatory perimeter map: for each action type, the regulatory regime(s) engaged (AI Act classification, MiFID II activity, AML obligation, GDPR Article 22 trigger), identifying any actions that approach a fiduciary boundary (ART-3).

A.4 Risk Profile

A.4 — ART and CRT Risk Assessment

(a) Applicable ART categories (ART-1 through ART-6) with severity assessment for this specific deployment, referencing the architecture profile from Section 5 where applicable.
(b) Applicable CRT categories (CRT-1 through CRT-8) from the CRSA-1 EU Edition with severity assessment.
(c) For each applicable risk category: the detection methodology deployed and the mitigation measures implemented.
(d) Residual risk disclosure: risks that remain after mitigation, communicated to the management body under DORA Article 5 and to deployers under AI Act Article 9(7).

A.5 Monitoring and Oversight

A.5 — Monitoring Infrastructure

(a) Three-tier monitoring specification: envelope compliance metrics, behavioural distribution metrics with control chart parameters, and outcome quality metrics.
(b) Safe halt specification: the complete halt sequence for this agent, including state reconciliation procedure and rollback capability per tool interaction.
(c) Logging specification: action chain log format, correlation identifier scheme, retention period, and causal reconstruction capability verification results.
(d) Mandatory review trigger configuration: the specific thresholds and conditions triggering mandatory review per Section 7.4.

B Agent-Specific Addendum to Article 25(4) Contract Terms

This addendum extends the Article 25(4) model contract terms published in the CRSA-1 EU Edition (Appendix B) with provisions specific to agentic AI systems. These terms address the tool provider relationship, not the foundation model provider relationship (which is covered by the EU Edition's base terms).

Clause AG.1 — Tool Capability Specification

The Tool Provider **SHALL** provide to the System Provider a Tool Capability Specification containing:
(a) the tool's function, input format, output format, error response format, and latency characteristics;
(b) known limitations on the tool's output accuracy, including conditions under which outputs degrade;
(c) data types the tool accesses or processes, including personal data categories where applicable; and
(d) known vulnerabilities relevant to agentic consumption, including susceptibility to injection through crafted queries.

Clause AG.2 — Agent-Specific Update Notification

The Tool Provider **SHALL** notify the System Provider no fewer than [specified number] business days before any change to: (a) the tool's output format or schema; (b) the tool's error response format; (c) the tool's latency characteristics; or (d) the data sources underlying the tool's responses. Changes to any of these elements may alter the agent's behaviour even without changes to the agent itself.

Clause AG.3 — Resilience Testing Participation

Where the tool supports a critical or important function, the Tool Provider **SHALL** participate in the System Provider's agent-specific resilience testing per DORA Article 30(3)(d), including:

(a) providing a test environment or sandbox that replicates production tool behaviour for agent testing purposes;
(b) cooperating with adversarial scenario testing, including injection resistance testing where the tester submits crafted queries through the agent to assess the tool's response; and
(c) cooperating with safe halt testing, including simulated tool unavailability during agent multi-step execution.

Clause AG.4 — Incident Cooperation for Agent Actions

In the event that an agent's interaction with the tool contributes to an ICT-related incident or serious incident, the Tool Provider **SHALL**:

(a) provide diagnostic information about the tool's behaviour during the incident period, including query logs, response logs, and any anomalies detected on the tool's side;
(b) cooperate with causal reconstruction (ART-6 mitigation) by providing timestamps, request identifiers, and response content for the agent's interactions during the incident window; and
(c) participate in root cause analysis within the DORA reporting timeline (final report within 1 month).

Clause AG.5 — Tool-Level Exit Provisions

The Tool Provider **SHALL** support the System Provider's exit strategy per DORA Article 30(3)(f) by:

(a) providing the tool's interface specification in sufficient detail to enable the System Provider to integrate an alternative tool implementing the same abstract interface;
(b) providing a transition period of no fewer than [specified number] months during which the current tool remains available while the System Provider validates the alternative; and
(c) cooperating with parallel running during the transition period, where both the current and alternative tools receive the same agent queries for output comparison.

C Agent Oversight Architecture Decision Matrix

This matrix maps specific financial services action types to the appropriate intervention model based on three criteria: reversibility, regulatory impact, and customer harm potential.

Action Type	Revers.	Reg. Impact	Harm Potential	Intervention Model
Trade execution (within limits)	Partial	MiFID II	Medium	Model 2 + 3
Trade execution (exceeding limits)	Partial	MiFID II	High	Model 1
Credit decision (approve)	Reversible	AI Act HR	Medium	Model 2 + 3
Credit decision (deny)	Irreversible*	AI Act HR	High	Model 1
Insurance claim approval (below threshold)	Partial	Solvency II	Low	Model 3
Insurance claim denial	Irreversible*	Solvency II	High	Model 1
Insurance claim (approaching pricing)	—	AI Act HR	High	Model 1
AML alert closure (low risk)	Reversible	AMLD	Medium	Model 2 + 3
AML alert closure (high risk)	Irreversible*	AMLD	High	Model 1
SAR filing / drafting	Irreversible	AMLD	High	Model 1 (always)
Account opening	Reversible	AMLD, GDPR	Low	Model 2 + 3
Account rejection	Irreversible*	GDPR Art. 22	High	Model 1
Customer communication	Partial	IDD, MiFID II	Medium	Model 2
Payment initiation	Time-bounded	PSD2	High	Model 1
Data source query (read-only)	Reversible	GDPR	Low	Model 3
External system write	Varies	DORA	High	Model 1

Table 5: Agent Oversight Architecture Decision Matrix. Irreversible indicates the action's effects on the affected person cannot be fully reversed even if the decision is subsequently overturned (reputational harm, credit record impact, delayed access to services). "AI Act HR" = AI Act high-risk classification. Intervention models: Model 1 (pre-commitment gate), Model 2 (statistical process control), Model 3 (post-action audit).*

The matrix reveals a consistent pattern: actions with irreversible effects on natural persons require Model 1 (pre-commitment gates) regardless of the agent's autonomy tier. This is the practical implementation of Article 14(4)(d)'s override requirement: for actions that cannot be meaningfully reversed, override must occur *before* execution.

D Glossary of Agentic AI Terms for Financial Services

Action Chain Log
A structured record of every action taken by an agent during a task execution, including the reasoning supporting each action, intermediate observations, tool selection rationale, and decision criteria at each branch point. Required for Article 14(4)(b) output interpretation and ART-6 causal reconstruction.

Action Envelope
The subset of the agent's capability envelope that the management body has approved for autonomous execution without per-action human confirmation. The boundary between the action envelope and the capability envelope defines the escalation boundary. See Section 2.2.

Agentic AI System
A composed AI system exhibiting all five defining characteristics: autonomous goal pursuit, dynamic tool selection, environmental observation, multi-step planning, and external state modification. See Section 1.2.

Agent Oversight Architecture
The three-model governance framework for autonomous agents: Model 1 (pre-commitment gates), Model 2 (statistical process control with automatic halt), Model 3 (post-action audit with reversal). Tier 3 agents require multiple concurrent models. See Section 3.6.

ART (Agentic Risk Taxonomy)
Six risk categories specific to autonomous agents in financial services: ART-1 Autonomous Decision Drift, ART-2 Tool Chain Escalation, ART-3 Fiduciary Boundary Violation, ART-4 Autonomous Concentration, ART-5 Cross-System State Propagation, ART-6 Accountability Void. See Section 4.

Attribution Window
The maximum lookback period (not exceeding 90 days) applied to determine the onset of a Category B (cumulative behavioural pattern) incident. See Section 7.2.

Autonomous Concentration (ART-4)
Emergent concentration risk arising when multiple agents independently converge on the same strategy, tool providers, or market positions without deliberate human direction. See Section 4.3.4.

Autonomous Decision Drift (ART-1)
Shift in an agent's decision boundary over time through accumulated context and environmental feedback, without a discrete model update. See Section 4.3.1.

Capability Envelope
The complete set of tools, data sources, and external systems the agent can access. Broader than the action envelope; the gap between them is the residual risk surface. See Section 2.2.

Cross-System State Propagation (ART-5)
Cascading effects in external systems triggered by the agent's actions but not observed, monitored, or controlled by the agent. See Section 4.3.5.

Fiduciary Boundary Violation (ART-3)
An agent's autonomous actions crossing a regulatory perimeter the institution did not anticipate, triggering obligations under a regulatory regime not addressed in the system's compliance framework. See Section 4.3.3.

Governance Threshold
The autonomy level at which existing regulatory frameworks exhibit material governance gaps. Threshold 1 (Tier 1 to Tier 2 boundary): gaps emerge. Threshold 2 (Tier 2 to Tier 3 boundary): existing frameworks break. See Section 1.3.

Portable Agent Specification (PAS)
A model-independent specification of an agent's behavioural contract, tool interfaces, action envelope, evaluation benchmarks, monitoring parameters, and regulatory compliance profile, designed to enable provider substitution under DORA exit strategy requirements. See Section 8.2.

Tool Chain Escalation (ART-2)
Emergent capability exceeding any individually assessed tool's scope, produced by the agent's dynamic composition of multiple tool invocations within a single task. See Section 4.3.2.

Tool Inventory Registration Methodology
The framework for mapping agent tool connections to the DORA Register of Information ITS template, with three approaches: provider-level, function-level, and capability-level registration. See Section 6.3.

Accountability Void (ART-6)
The absence of a unified causal model connecting distributed agent logs into a single accountable narrative when adverse outcomes arise from multi-agent delegation chains. See Section 4.3.6.

Tier 1 — Assistive Agent
An agent where the human approves every externally visible action. Existing governance frameworks are adequate. See Section 1.3.

Tier 2 — Supervised Autonomous Agent
An agent operating autonomously within a pre-approved action envelope, with exception escalation to human oversight. Governance gaps emerge. See Section 1.3.

Tier 3 — Fully Autonomous Agent
An agent completing tasks end-to-end including material external state modification without per-action human approval. Existing governance frameworks break. See Section 1.3.

References

Primary EU Legislation

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *OJ L, 2024/1689, 12.7.2024.*

[2] Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector (DORA). *OJ L 333, 27.12.2022.*

[3] Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions (CRR). *OJ L 176, 27.6.2013.*

[4] Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments (MiFID II). *OJ L 173, 12.6.2014.*

[5] Directive (EU) 2015/849 of the European Parliament and of the Council of 20 May 2015 on the prevention of the use of the financial system for the purposes of money laundering or terrorist financing (AMLD4). *OJ L 141, 5.6.2015.*

[6] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data (GDPR). *OJ L 119, 4.5.2016.*

[7] Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II). *OJ L 335, 17.12.2009.*

[8] Directive (EU) 2016/97 of the European Parliament and of the Council of 20 January 2016 on insurance distribution (IDD). *OJ L 26, 2.2.2016.*

[9] Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products (Product Liability Directive). *OJ L, 2024/2853, 18.11.2024.*

DORA Regulatory Technical Standards and Implementing Technical Standards

[10] Commission Delegated Regulation (EU) 2024/1774 of 13 March 2024 supplementing Regulation (EU) 2022/2554 with regard to regulatory technical standards specifying ICT risk management tools, methods, processes, and policies. *OJ L, 2024/1774.*

[11] Commission Delegated Regulation (EU) 2024/1772 of 13 March 2024 supplementing Regulation (EU) 2022/2554 with regard to regulatory technical standards specifying the criteria for the classification of ICT-related incidents. *OJ L, 2024/1772.*

[12] Commission Implementing Regulation (EU) 2024/2956 of 2 December 2024 laying down implementing technical standards for the application of Regulation (EU) 2022/2554 with regard to the standard templates for the Register of Information. *OJ L, 2024/2956.*

[13] Commission Delegated Regulation (EU) 2025/301 supplementing Regulation (EU) 2022/2554 with regard to regulatory technical standards specifying the content and time limits for incident reporting. *OJ L, 2025/301.*

[14] Commission Delegated Regulation (EU) 2025/1190 supplementing Regulation (EU) 2022/2554 with regard to regulatory technical standards specifying the criteria for threat-led penetration testing. *OJ L, 2025/1190.*

[15] Commission Delegated Regulation (EU) 2025/532 supplementing Regulation (EU) 2022/2554 with regard to regulatory technical standards specifying the elements for assessing sub-outsourcing risks. *OJ L, 2025/532.*

[16] Commission Delegated Regulation (EU) 2024/1502 supplementing Regulation (EU) 2022/2554 with regard to the criteria for designation of critical ICT third-party service providers. *OJ L, 2024/1502.*

Commission Proposals and Guidance

[17] European Commission. Proposal for a Regulation amending Regulations (EU) 2024/1689 and (EU) 2024/2847 (Digital Omnibus on AI). COM(2025) 836 final, 6 February 2025.
[18] European Commission. Commission Guidelines on the definition of an artificial intelligence system (C(2025) 3034 final), 4 February 2025.
[19] European Commission. Draft guidance on reporting of serious incidents under Article 73 of Regulation (EU) 2024/1689. Published for consultation, September 2025.
[20] European Commission. Clarification on AI systems used for the purpose of detecting financial fraud under Annex III point 5(b). May 2025.

Supervisory Guidance and Publications

[21] BaFin. "Guidance on ICT Risks in the Use of AI at Financial Entities." 18 December 2025. 35 pp.
[22] European Central Bank. Revised Guide to Internal Models (EGIM). 25 July 2025. Including new chapter on machine learning techniques.
[23] European Central Bank. Supervisory Priorities 2026–2028. November 2025.
[24] European Central Bank. "Technology is neutral, governance is not: AI adoption in the banking sector." Speech, 24 February 2026.
[25] European Central Bank. "AI's impact on banking: use cases for credit scoring and fraud detection." Supervisory Newsletter, November 2025.
[26] European Banking Authority. "AI Act implications for the EU banking sector." Factsheet, November 2025.
[27] European Banking Authority. Guidelines on loan origination and monitoring. EBA/GL/2020/06. 29 May 2020.
[28] European Banking Authority. Discussion Paper on machine learning for IRB models. EBA/DP/2021/04. November 2021.
[29] European Banking Authority. Follow-up report on the use of machine learning for internal ratings-based models. EBA/REP/2023/28. August 2023.
[30] European Banking Authority. Amendment to Guidelines on ICT and security risk management (EBA/GL/2019/04). 11 February 2025.
[31] EIOPA. Opinion on AI governance and risk management in the insurance sector. 6 August 2025.
[32] ESA Joint Committee. Designation of critical ICT third-party service providers under DORA. 18 November 2025.
[33] European Data Protection Board. Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679. WP251rev.01, as last revised February 2018.

International Standards and Publications

[34] Federal Reserve Board / Office of the Comptroller of the Currency. SR 11-7 / OCC 2011-12: Supervisory Guidance on Model Risk Management. 4 April 2011.
[35] Basel Committee on Banking Supervision. Newsletter on artificial intelligence and machine learning. March 2022.
[36] Basel Committee on Banking Supervision. Report on digitalisation of finance. May 2024.
[37] BIS Financial Stability Institute. Occasional Paper No. 24: AI explainability regulation. 2025.
[38] Financial Stability Board. Financial stability implications of artificial intelligence. November 2024.

[39] Financial Stability Board. Follow-up report: monitoring AI-related third-party dependencies. October 2025.

Industry and Academic Sources

[40] Harvard Data Science Review. "The Future of Credit Underwriting and Insurance Under the EU AI Act." Issue 7.3, Summer 2025.
[41] Microsoft. "Managing concentration risk and exit requirements: A framework for financial institutions." Industry Blog, 2 February 2026.
[42] Pinsent Masons. "Financial services compliance with the EU AI Act and DORA can be streamlined." Out-Law Analysis, 2025.
[43] KPMG. "ECB's New Perspective on Machine Learning in Banking." 2025.

Auburn Governance Stack

[44] Fields, R. "CRSA-1 EU Edition: Compositional Runtime Safety Attestation Protocol — Compositional Safety Compliance Profile for EU AI Act Regulation 2024/1689." Auburn Governance Stack, 2026.
[45] Fields, R. "CRSA-1: Compositional Runtime Safety Attestation Protocol for Multi-Principal AI Systems." Auburn Governance Stack, 2026.
[46] Fields, R. "The Model Attestation Interface (MAI-1): A Normative Profile and Conformance Protocol for Foundation Model Governance." Auburn Governance Stack, 2026.
[47] Fields, R. "CTS-1: MAI-1 Conformance Test Suite." Auburn Governance Stack, 2026.
[48] Fields, R. "Auburn Governance Stack: Master Architecture Plan." Auburn Governance Stack, 2026.

Intellectual Property Declaration

The methods, logic structures, Agentic Risk Taxonomy (ART-1 through ART-6), Agent Oversight Architecture (three intervention models), autonomy tier classification framework, Tool Inventory Registration Methodology, Agent Incident Classification Decision Logic, Portable Agent Specification, Agent Capability Declaration Template, and agent-specific contract terms contained in this work are the sole property of Ryan Fields.

Public License (Non-Commercial)

- **Academic Use:** Researchers may share and use this framework for non-commercial academic purposes, provided full attribution is given to Ryan Fields.
- **No Derivatives:** No modifications or adaptations of the Agentic Risk Taxonomy, Agent Oversight Architecture, autonomy tier classification, incident classification decision logic, Portable Agent Specification, or governance methodologies are permitted without express written consent.

Commercial Prohibition

Commercial use of this framework is strictly prohibited. This includes, but is not limited to:

- Use within proprietary AI governance, risk management, or compliance software.
- Integration into commercial model risk management platforms, RegTech solutions, or supervisory technology tools.
- Use by consulting firms, law firms, auditors, or advisory practices in client-facing deliverables.
- Incorporation into financial institution compliance documentation, internal governance frameworks, or regulatory filings without license.
- Use by notified bodies, conformity assessment bodies, or supervisory authorities in assessment methodologies without license.

Ryan Fields
UncleBroFields@proton.me
fieldsryanchristopher@gmail.com

Autonomous AI Agents in Regulated Financial Services
CRSA-1 EU Compliance Series
Fields, 2026