AI Guardian

Doc type: technical · **Version:** v0.1 · **Status:** published · **Module slug:** ai-guardian
Exported: 2026-05-15 11:12 UTC · **By:** anonymous

AI Guardian – Technical Specification

1. Module Purpose & Scope (Authoritative)

AI Guardian is an optional Intelligence Suite module that acts as a 24/7 operational analyst for Primoro. Its purpose is to continuously audit workflow completeness, diary integrity, and operational consistency, and to convert any detected gap or risk into a clear, owned next step for the team. AI Guardian supports people — it never replaces judgement, authority, or ownership.

It governs:

- Continuous auditing of derived operational signals across Primoro modules and conversion of detected gaps or risks into governed Findings.
- Routing of actionable outputs (tasks, alerts, summaries) to the correct role via Task Manager or Communication Hub.
- Enforcing closed-loop resolution: finding → action → resolved, with a full immutable audit trail.

It explicitly does not:

- Monitor or score individual staff performance — no module in the current corpus owns this concern; it is out of scope for the platform.
- Act as a system of record for any primary entity — that responsibility belongs to the originating module (e.g. Appointment Manager for appointments, Task Manager for tasks).
- Override CORE rules governing appointments, deposits, access, or eligibility — those boundaries are owned by their respective modules.

---

2. Ownership & Responsibilities

2.1 AI Guardian IS Responsible For

- Auditing derived operational signals originating in connected modules and detecting missed actions, incomplete workflows, unsafe states, and SLA risk.
- Creating and lifecycle-managing Guardian Findings as the canonical governed artefact for every detected issue.
- Routing all actionable outputs exclusively through Task Manager (tasks) or Communication Hub (alerts and summaries).
- Enforcing closed-loop resolution so that no Finding can be cleared without an auditable resolution event.
- Maintaining an immutable audit trail for every Finding, including source signals, all state transitions, escalation events, and resolution identity.

2.2 AI Guardian IS NOT Responsible For

- Staff performance monitoring, ranking, grading, or comparison — this is an explicit out-of-scope concern; no current module owns it.
- Taking autonomous actions without explicit human confirmation — human approval is required for every committed output.
- Overriding governance, access control, or eligibility decisions — owned by Access Manager and the relevant CORE modules.
- Consuming raw financial transaction data — Financial Insights owns that surface; AI Guardian consumes aggregated operational signals only.
- Producing quality findings or remediation tasks — owned by AI Quality Monitor; AI Guardian ingests those outputs as signals and ensures governed follow-through.

---

3. Core Objects (Normative)

3.1 Guardian Finding (Canonical Artefact)

A Guardian Finding is a governed digital artefact representing one detected operational gap, risk, or inconsistency that requires a human-owned next step.

Minimum required fields:

- FindingID (unique)
- SourceSignals (one or more originating signals)
- FindingType (e.g. missed action, SLA risk, inconsistency)
- Severity (informational / warning / critical)
- LinkedEntity (appointment, task, patient, diary day, etc.)
- FindingState
- OwningRole
- CreatedAt
- AuditTrail (immutable)

Findings are immutable in origin; only their state may change.

3.2 Guardian Finding State Machine (Authoritative)

States:

- Detected
- Action Created
- In Progress
- Escalated *(optional)*
- Resolved
- Closed

Rules:

- A Finding MUST NOT remain in **Detected** without an associated action being created.
- Resolution requires evidence: either task completion or an explicit, audited dismissal with a stated reason.
- Closing a Finding clears all related alerts.
- State transitions are auditable and time-stamped; no transition may be applied without an identified actor recorded in the audit trail.
- A Finding MUST NOT return to Detected once it has reached Action Created or beyond.

---

4. Detection & Signal Consumption

4.1 System-Initiated Detection (Authoritative)

AI Guardian detects Findings from derived signals originating in:

- **Appointment Manager** — booking, completion, and cancellation events.
- **Task Manager** — open, overdue, and missing task states.
- **Communication Hub** — delivery status, acknowledgements, and SLA timers.
- **Payments / deposits** — state signals only; card data is never consumed.
- Recall and follow-up logic.
- **Financial Insights** — revenue anomalies, deposit state irregularities, subscription status changes, and provider production outliers. Signal scope is limited to aggregated operational events; no raw financial transactions are consumed directly.
- **Aftercare Manager** — higher-risk patient patterns detected during aftercare workflows. When AI Guardian is enabled, findings from this domain are prioritised for earlier review.
- **AI Quality Monitor** — quality findings, remediation tasks, and exception outputs that require Guardian-level follow-through. AI Guardian ingests these as operational signals and ensures each results in a governed Finding with an owned action.

The module MUST:

- Convert every detected signal that crosses a configured threshold into exactly one Guardian Finding.
- Audit the source signal(s) against the resulting Finding at creation time.

The module MUST NOT:

- Consume raw financial transaction records (owned by Financial Insights).
- Block or delay live workflows while performing continuous audit.

4.2 AI Quality Monitor Signal Alignment

When AI Guardian consumes signals from AI Quality Monitor, it MUST align with AI Quality Monitor's own Quality Finding lifecycle. Specifically:

- AI Guardian treats a Quality Finding as an inbound operational signal at the point AI Quality Monitor transitions it to a state that requires external follow-through (e.g. a remediation task has been raised and not acted upon, or an exception has been escalated beyond Quality Monitor's own resolution boundary).
- The attribution model and evidence clips associated with a Quality Finding are propagated as source-signal metadata on the resulting Guardian Finding, so that AI Guardian's audit trail can reference the originating

evidence without duplicating or re-owning it.

- AI Guardian MUST NOT alter, close, or dismiss a Quality Finding directly; any governed action on the Quality Monitor artefact remains the responsibility of AI Quality Monitor. Guardian's role is to ensure a human-owned next step exists and is tracked to resolution.
- Where a Quality Finding and a Guardian Finding coexist for the same underlying event, the Guardian Finding MUST carry a reference to the originating Quality FindingID to support holistic audit reconstruction.

4.3 Integrated Payments & Subscription State Signals

Integrated Payments exposes all payment lifecycle and subscription states as read-only signals to AI Guardian. These signals supplement the deposit and payment state signals already listed in §4.1 and extend coverage to:

- **Payment lifecycle events** — initiated, authorised, failed, refunded, and disputed payment states.
- **Subscription state changes** — new subscriptions, renewals, lapses, cancellations, and reinstatements.
- **Instalment plan anomalies** — missed instalments and plan state deviations that cross a configured threshold.

AI Guardian MUST consume these signals exclusively as aggregated operational events via the Integrated Payments integration contract; no raw card or transaction data is consumed. A Guardian Finding MAY be raised where a subscription lapse or payment anomaly is assessed as creating an operational or compliance risk for the practice.

4.4 Communication Hub Governance Audit Signals

Communication Hub is governed by a core principle of replacing ungoverned, ad-hoc communication with auditable, policy-bound channels. AI Guardian MUST therefore ingest Communication Hub governance audit events as a distinct signal category, covering:

- **Blocked group-chat attempts** — where a staff member has attempted to initiate or continue an ungoverned group communication and the attempt has been blocked by Communication Hub policy.
- **Ungoverned communication attempts** — repeated or persistent attempts to route patient or clinical communication outside governed channels.
- **Policy violation events** — any event flagged by Communication Hub as a breach of its configured communication governance rules.
- **SLA violations** — delivery or acknowledgement SLA breaches that Communication Hub surfaces for operational review.

Where patterns of attempted circumvention (e.g. repeated blocks within a configurable time window) cross a configured threshold, AI Guardian MUST raise a Guardian Finding for management review. Individual one-off blocks below threshold are consumed as signals but do not automatically produce Findings. The detection threshold for this signal category is configurable per practice via the Admin Control Plane (see §13.3).

4.5 User-Initiated Actions (Authoritative)

- Managers MAY dismiss or resolve a Finding with a stated reason; this action is fully audited.
- No Finding may be dismissed or resolved by a role below Manager without an explicit escalation or approval step, as governed by Access Manager RBAC.

---

5. Delivery Surfaces & Access (Authoritative)

5.1 Web Portal

The web portal MUST provide:

- A Guardian Findings list, filterable by severity, status, and role.
- A Finding detail panel showing source signals, AI reasoning, linked entity, and full audit history.
- Inline task linkage to related Task Manager records.
- Explicit **Resolve**, **Escalate**, and **Dismiss** actions, each RBAC-controlled.

5.2 Tablet App

- Read-only notifications for critical Findings only.
- Resolution and dismissal actions are not available on the tablet surface.

5.3 Patient Mobile App

AI Guardian has no patient-facing surface. No Guardian Finding or underlying signal is exposed directly to patients.

5.4 Engagement Signals

- AI Guardian emits Finding counts by severity and state to staff dashboards for operational visibility.
- SLA breach records, where configured, are surfaced as engagement signals for practice managers.

---

6. Integration Contracts

6.1 Inbound (this module consumes from)

From module	What	Contract
Appointment Manager	Booking, completion, cancellation events	async event
Task Manager	Open, overdue, missing task states	async event
Communication Hub	Delivery status, acknowledgements, SLA timers, governance audit events (blocked communications, policy violations)	async event
Payments / Integrated Payments	Deposit, payment lifecycle, and subscription state signals (no card data)	async event

Financial Insights	Aggregated operational signals (anomalies, outliers)	async event
Aftercare Manager	Higher-risk patient pattern signals	async event
AI Quality Monitor	Quality findings, remediation tasks, exception outputs	async event

6.2 Outbound (this module emits to)

To module	What	Contract
Task Manager	Structured tasks for human follow-up on Findings	event
Communication Hub	Alerts and summaries for staff notification	event
Audit & Compliance	Immutable Finding audit log entries	event
Security & Privacy	SecurityEvent references for security- or compliance-material Findings	event

6.3 PMS Boundary

AI Guardian does not read from or write to the PMS directly. All signals that originate in PMS-adjacent data (e.g. appointment records) reach AI Guardian exclusively through the Appointment Manager integration contract; AI Guardian does not hold a direct PMS integration.

6.4 SecurityEvent Linkage

Guardian Findings that are material to security compliance or incident investigation MUST emit a corresponding SecurityEvent reference to the Security & Privacy module via the outbound contract in §6.2. This ensures that compliance audit trails (GDPR, NHS DSPT, and inspection requirements) can be reconstructed holistically across the platform without AI Guardian duplicating the SecurityEvent record. Specifically:

- AI Guardian MUST NOT own or author SecurityEvent records; it emits a reference (FindingID, FindingType, severity, timestamp, and linked entity) that Security & Privacy uses to associate the Guardian Finding with its own canonical SecurityEvent stream.
- The threshold for emitting a SecurityEvent reference is: any Guardian Finding of severity **warning** or **critical** whose FindingType relates to access anomalies, policy violations, ungoverned communication patterns, or data handling irregularities.
- Informational Findings do not require SecurityEvent linkage unless explicitly escalated to warning or critical.

---

7. AI Boundaries (Non-Negotiable)

AI MAY:

- Detect operational gaps and risks from derived signals and surface them as governed Findings for human review.
- Generate reasoning text explaining why a Finding was raised, linked to source signals.
- Suggest tasks or alert content for human approval before those outputs are committed to Task Manager or Communication Hub.
- Summarise Finding activity for staff dashboards.

AI MAY NOT:

- Resolve, dismiss, or close a Finding autonomously — explicit human action is required in every case.
- Rank, score, grade, or compare individual staff members.
- Override policy-bound decisions in appointments, access, deposits, or eligibility.
- Bypass governance, audit, or access control checks.
- Make commitments on behalf of the practice.
- Take any action whose reasoning cannot be surfaced and explained to a human reviewer.

7.1 Boundary with AI Assistant (Aiden)

AI Guardian and AI Assistant (Aiden) are separate Intelligence Suite modules with complementary but non-overlapping roles. Their boundaries are as follows:

- **Aiden** operates as a conversational assistant and applies escalation rules and clinical safety checks at the point of individual staff or patient interaction. Aiden's safety boundaries govern what it will and will not surface or suggest in a single conversational turn.
- **AI Guardian** operates as a continuous background audit process across all modules. It does not participate in conversational interactions and does not intercept or modify Aiden's outputs. Guardian's escalation rules apply to the lifecycle of governed Findings, not to individual conversational exchanges.
- Where Aiden raises a safety-related escalation event (e.g. a confidence threshold not met, or a clinically sensitive query routed for human review), that escalation event MAY be surfaced to AI Guardian as an operational signal if it meets a configured detection threshold — resulting in a Guardian Finding for management visibility. Aiden does not receive Guardian Findings directly.
- Audit events from Aiden and audit events from AI Guardian are both routed to Audit & Compliance but remain distinct record types; neither module consumes the other's audit stream for detection purposes.
- No Guardian AI boundary may be relaxed on the basis that Aiden has already applied a safety check; each module enforces its own boundaries independently.

---

8. Audit & Compliance

The system MUST log (immutable):

- Finding ID and type at creation.
- Source signal(s) linked to the Finding at creation time.
- Timestamps for all state transitions (Detected → Action Created → In Progress → Escalated → Resolved → Closed).
- Owning role assigned at each state.

- All tasks created as outputs of a Finding, with linkage to Task Manager records.
- All escalation events, including escalating actor and timestamp.
- Resolution method (task completion or explicit dismissal) and resolver identity.
- All dismissal reasons provided by managers.
- All AI suggestions surfaced to staff, including which were accepted and which were rejected, with timestamps and actor identities.
- SecurityEvent reference IDs emitted to Security & Privacy for compliance-material Findings (see §6.4).

Audit logs MUST be immutable and exportable for inspection. All actions must be attributable and inspection-ready.

---

9. Access Control

Access control for AI Guardian is governed by Access Manager RBAC. The following capabilities are role-gated:

- **View Findings list and detail** — staff roles as configured per practice.
- **Create tasks or alerts from a Finding** — manager role or above.
- **Resolve or dismiss a Finding** — manager role or above; reason is mandatory and audited.
- **Escalate a Finding** — manager role or above.
- **Configure Guardian thresholds and signal scope** — practice admin via Admin Control Plane.
- Deletion of Findings is not permitted; the audit trail is immutable.

MFA requirements for sensitive operations (e.g. bulk dismissal of critical Findings) are governed by Access Manager policy and apply where Access Manager mandates them.

---

10. Integration Summary

- **Appointment Manager** — inbound async events: booking, completion, cancellation signals.
- **Task Manager** — inbound async: open/overdue task signals; outbound event: structured tasks for Finding follow-up.
- **Communication Hub** — inbound async: delivery, SLA, and governance audit signals (blocked communications, policy violations); outbound event: alerts and summaries.
- **Financial Insights** — inbound async: aggregated operational signals (anomalies, outliers).
- **Integrated Payments** — inbound async: payment lifecycle and subscription state signals.
- **Aftercare Manager** — inbound async: higher-risk patient pattern signals.
- **AI Quality Monitor** — inbound async: quality findings and remediation task signals.
- **Access Manager** — RBAC enforcement for all resolve, dismiss, escalate, and configure actions.
- **Audit & Compliance** — outbound event: immutable Finding audit log.
- **Security & Privacy** — outbound event: SecurityEvent references for compliance-material Findings.

---

11. Explicit Non-Goals

- **Staff performance monitoring, ranking, or grading** — explicitly prohibited; no current module owns this concern and it is out of scope for the platform.
- **Autonomous resolution of Findings** — human confirmation is required for every committed action; autonomous resolution is architecturally prohibited.
- **Direct PMS integration** — AI Guardian reads no PMS data directly; all signals arrive through owning modules.
- **Raw financial transaction consumption** — aggregated signals via Financial Insights only.
- **Patient-facing surfaces** — AI Guardian has no patient-visible output.
- **Authoring SecurityEvent records** — AI Guardian emits SecurityEvent references only; Security & Privacy owns the canonical SecurityEvent stream.

---

12. Versioning & Governance

This specification is owned by: *(Intelligence Suite module owner — role to be confirmed)*

Changes to this spec require:

- Review by the Post-MVP module owner.
- Impact analysis across all declared related modules (see /propose).
- Version bump (patch / minor / major) depending on scope of change.

---

13. Build Contract (Engineering & QA)

13.1 Canonical Data Model

(no content captured in original — needs definition)

The canonical artefact is the **Guardian Finding**. Minimum fields are defined in §3.1. Full schema — including index strategy, foreign key constraints, and soft-delete policy — requires engineering definition before build.

13.2 Core Behaviour Rules

The following rules are testable and must be implemented by engineering and verified by QA:

1. Every detected signal that crosses a configured threshold MUST produce exactly one Guardian Finding; duplicate Findings for the same signal instance are not permitted.
2. A Finding MUST NOT remain in **Detected** state without an associated action (task or alert) being created.
3. No Finding may be resolved or closed without an auditable resolution event (task completion or explicit manager dismissal with reason).
4. No Guardian output may rank, score, grade, or compare individual staff members.
5. All Finding outputs MUST route exclusively through Task Manager (tasks) or Communication Hub (alerts and summaries); no direct notification channel is permitted.
6. All state transitions on a Finding MUST be time-stamped and actor-attributed in the immutable audit trail.
7. AI-generated reasoning text MUST be surfaced alongside each Finding and linked to the source signal(s) that triggered it.

8. AI Guardian MUST NOT consume raw financial transaction data; only aggregated operational signals from Financial Insights are permitted.

9. Tablet surfaces MUST be read-only and MUST NOT expose resolve, dismiss, or escalate actions.

10. Closing a Finding MUST clear all related alerts emitted via Communication Hub.

11. A Guardian Finding derived from an AI Quality Monitor signal MUST carry a reference to the originating Quality FindingID in its AuditTrail; AI Guardian MUST NOT close or dismiss the Quality Finding directly.

12. A Guardian Finding of severity warning or critical whose FindingType relates to access anomalies, policy violations, ungoverned communication patterns, or data handling irregularities MUST emit a SecurityEvent reference to Security & Privacy.

13. Communication Hub governance audit events (blocked communications, policy violations) MUST be evaluated against configured thresholds; patterns crossing the threshold MUST produce a Guardian Finding.

14. Payment lifecycle and subscription state signals from Integrated Payments MUST be evaluated as operational signals; Guardian Findings MAY be raised where anomalies cross a configured operational risk threshold.

13.3 Configuration Surfaces

- **Practice-level settings** (Admin Control Plane): enable/disable AI Guardian, configure detection thresholds per signal type (including Communication Hub governance audit signal thresholds and Integrated Payments anomaly thresholds), configure severity mappings.
- **Per-role settings** (Access Manager): which roles receive which severity of notification.
- **Per-Finding overrides**: managers may escalate or adjust owning role assignment on an individual Finding (audited).

13.4 Filtering & Views

The web portal Findings list MUST support filtering by:

- Severity (informational / warning / critical)
- Finding state (Detected / Action Created / In Progress / Escalated / Resolved / Closed)
- Owning role
- Source module (e.g. Appointment Manager, Task Manager, AI Quality Monitor, Communication Hub, Integrated Payments)
- Date range of detection

Saved views are configurable per user via Access Manager preferences.

13.5 Module Extension Map

- Additional signal sources MAY be added by declaring a new inbound integration contract in §6.1 without altering the core Finding state machine.
- New Finding types MAY be introduced as enumerated values in the FindingType field; no schema migration to existing Findings is required.
- The severity scale (informational / warning / critical) is the stable contract surface; introducing new severity levels is a minor version change requiring impact analysis.

13.6 Acceptance Criteria

The build of AI Guardian is complete when:

- [ ] Every detected issue produces exactly one governed Finding with all required fields populated.
- [ ] Finding state machine transitions enforce all rules in §3.2 and §13.2.
- [ ] A Finding cannot remain in Detected without an action.
- [ ] No Finding can be resolved or closed without an auditable resolution event.
- [ ] All outputs route exclusively through Task Manager or Communication Hub.
- [ ] No Guardian output ranks, scores, grades, or compares individuals.
- [ ] All integrations in §6 are wired and verified with test events.
- [ ] AI boundaries in §7 are enforced; negative tests (e.g. autonomous resolution attempts) pass.
- [ ] Aiden/Guardian boundary rules in §7.1 are enforced; neither module processes the other's audit stream for detection.
- [ ] Audit log captures every event in §8, is immutable, and is exportable.
- [ ] Access control per §9 is enforced; role-gating negative tests pass.
- [ ] Tablet surface is read-only with no resolve/dismiss/escalate actions present.
- [ ] Guardian Findings derived from AI Quality Monitor signals carry originating Quality FindingID references and do not alter the source Quality Finding.
- [ ] SecurityEvent references are emitted to Security & Privacy for all qualifying Findings (rule 12 in §13.2).
- [ ] Communication Hub governance audit signal thresholds are configurable and trigger Findings correctly.
- [ ] Integrated Payments subscription and payment lifecycle signals are consumed and evaluated against configured thresholds.
- [ ] All non-functional requirements in §14 are met.

---

14. Non-Functional Requirements

- **Performance:** AI Guardian's continuous audit process MUST NOT block or measurably delay live workflows in any integrated module. Finding detection and routing latency targets require definition before build (captured in §15).
- **Reliability:** Findings persist until explicitly resolved or dismissed; no Finding may be silently lost. The module should degrade gracefully if a signal source is temporarily unavailable — queuing inbound events rather than dropping them — with availability targets to be defined before build.
- **Scalability:** The module must support multi-site, multi-practice deployments without cross-tenant data leakage; signal processing must scale with the number of connected modules and configured detection rules.
- **Security:** Least-privilege access enforced via Access Manager. All patient-bound data consumed as signals must be encrypted in transit and at rest. Secrets and credentials for integration surfaces must follow platform key management standards.
- **Privacy:** AI Guardian consumes patient-linked signals and must honour applicable GDPR rights (access, erasure where technically compatible with immutable audit requirements). Data retention policy for Guardian Findings requires definition before build (captured in §15).
- **Observability:** The module must export metrics covering Finding detection rate, mean time to action, mean time to resolution, and escalation rate. Structured logs must be emitted for all state transitions. Distributed tracing must cover the signal-ingestion-to-Finding-creation path.

---

15. Open Questions

1. **Detection latency targets:** The original spec states that continuous audit must not block live workflows, but does not define acceptable end-to-end latency from signal emission to Finding creation. What are the target SLAs?
2. **Availability target:** No uptime or graceful-degradation SLA is defined. What is the required availability for the Guardian audit process, and what is the acceptable behaviour when a signal source module is unavailable?
3. **Data retention policy for Findings:** How long must resolved and closed Findings be retained in the system? Does this differ by Finding severity or linked entity type?
4. **GDPR erasure and immutable audit tension:** Where a patient exercises a right-to-erasure request, how should AI Guardian handle Findings whose audit trail references that patient? The immutability requirement and the erasure right are in tension — a resolution approach needs to be defined.
5. **Configurable thresholds:** The spec states that AI Guardian detects signals that cross a configured threshold, but does not define who sets thresholds, what the defaults are, or whether thresholds can be set per signal type. These defaults require product definition.
6. **Escalated state trigger:** The Escalated state is marked optional in the state machine. The conditions under which a Finding is escalated, and who may trigger escalation, are not fully defined.
7. **Recall and follow-up logic as a signal source:** This is listed as a source in §5.1 of the original but is not attributed to a named module. Which module owns recall and follow-up logic, and what is the formal integration contract?
8. **AI Quality Monitor Finding state alignment:** AI Guardian consumes Quality Findings at the point they require external follow-through, but the exact Quality Monitor states that trigger Guardian ingestion have not been formally agreed between the two modules. These trigger states require cross-module definition before build.
9. **Aiden escalation event threshold:** Where Aiden raises a safety-related escalation event that may be surfaced to AI Guardian as an operational signal (§7.1), the configurable threshold governing when this produces a Guardian Finding requires product definition.