Governing Local-First Agentic AI for Vulnerable Populations

A Balian AI White Paper

EXECUTIVE SUMMARY

Artificial Intelligence (AI) agents offer unprecedented opportunities to support vulnerable populations across crisis triage, assistive communication, and public sector casework. However, deploying these capabilities via cloud-hosted frontier models introduces systemic risks: *data exfiltration, regulatory non-compliance, and the economic precarity of recurring token costs*. Furthermore, traditional probabilistic AI guardrails cannot deterministically enforce the strict statutory boundaries required in high-stakes human services.

This white paper establishes that safely leveraging AI for vulnerable cohorts requires a fundamental paradigm shift from public, multi-tenant SaaS platforms to "private-core" computing—hosted either locally on-premises or within a secure, dedicated cloud tenancy. To operationalise this shift, *Balian AI*—a Singapore based applied AI practice specialising in defensible, enterprise-grade deployments—has designed a *9-Layer governed architecture*. By adopting this stack, organisations can bring capable open-weight models inside their own physical or virtual walls, ensuring absolute data sovereignty and offline operational resilience

Crucially, this architecture operationalises the Safeguards for Agentic Finance at Runtime (SAFR) guidance framework to decouple an agent's reasoning from its authority to act. Every proposed AI intervention is routed through a *deterministic* runtime interceptor that evaluates the action against explicit, machine-readable policy mandates. Actions are instantly categorised to Auto-Execute, Observe, Escalate to a human maker-checker, or Deny, with every step immutably recorded to a cryptographic hash-chain ledger.

Capable models are no longer the primary bottleneck in social care; the true barrier is operational discipline and defensible governance. Organisations cannot rely on commercial cloud vendors to *police trauma-informed data* through opaque prompt filters. By adopting a structurally governed, local-first architecture, institutions can guarantee that no consequential action executes without pre-execution validation, ensuring that AI deployments remain humane, safe, and entirely accountable to human professionals and regulatory bodies.

1. The Vulnerability Paradox in Agentic AI

The application of AI in social care and human services creates a *"vulnerability paradox":* the cohorts most in need of AI-enabled assistance are the least positioned to absorb the risks of its failure or exploitation. Existing cloud-reliant AI deployments amplify these risks across three primary vectors:

- **Data Sovereignty and Surveillance Exposure:** Transmitting domestic violence shelter locations, paediatric casework files, or neurodivergent behavioural profiles to public, multi-tenant cloud APIs compromises data sovereignty. It subjects highly sensitive information to third-party vendors' terms of service, secondary data monetisation, and the risk of catastrophic data breaches. For individuals in crisis, data leakage across public networks creates immediate physical safety hazards. While hosting models within a dedicated private cloud tenancy successfully prevent data leakage—as the cloud provider cannot read encrypted data—the high infrastructure costs often put this option out of reach for non-governmental organisations (NGOs) and SMEs.
- **Economic and Connectivity Precarity:** Cloud models rely on recurring token consumption,

Copyright © Shasana Balian AI Pte Ltd 2026

creating unsustainable financial burdens for underfunded non-governmental organisations (NGOs). Additionally, cloud-tethered assistive technologies fail in bandwidth-constrained rural, Indigenous, or disaster-response environments, creating dangerous operational blackouts.

- **Failure Modes of Purely Probabilistic Guardrails:** Traditional AI deployments apply static, probabilistic guardrails solely to initial prompts and final responses. While effective for general content moderation, probabilistic filters cannot deterministically guarantee compliance with hard statutory thresholds (e.g., executing emergency cash disbursements or statutory child protection interventions).

The greatest risk in these environments emerges when *human reviewers quietly stop checking*. As model performance appears reliable during routine tasks, caseworkers may fall victim to "automation bias," causing the critical human expertise that made oversight meaningful to atrophy. Without structural, systemic checks, the safety net fails.

2. Local-First Computing as an Enabling Technology

Local-first computing—running capable open-weight models within an organisation's physical or virtual walls—removes dependence on external cloud providers. This approach serves not only as a structural safeguard but also as a fundamental enabler of accessibility and inclusion.

- **Offline Resilience and Sovereignty:** Local-first architectures ensure continuous, uninterrupted service for remote Indigenous communities, regional care facilities, and mobile crisis responders who operate without reliable internet connectivity.
- **Predictable Fixed-Cost Inference:** By shifting to local-weight inference on edge hardware or private cloud tenancies, organisations eliminate variable SaaS token metering. While running independent infrastructure carries upfront capital expenditure and ongoing operational costs, it replaces unpredictable API pricing with a stable cost structure.
- **Absolute Privacy Isolation:** A private core ensures that sensitive trauma histories and medical evaluations never traverse the public internet, completely neutralising the threat of third-party corporate surveillance or secondary analytics profiling.

3. The 9-Layer Architecture: Fail-Closed Governance

To operationalise local-first AI safely, institutions must adopt a governed architecture that strictly decouples an agent's reasoning capabilities from its authority to act. The Enterprise AI Architecture Stack provides a 9-Layer (L1–L9) framework partitioned across three distinct security perimeters: Public Edge (Perimeter 1), Secure Bridge (Perimeter 2), and Private Core (Perimeter 3).

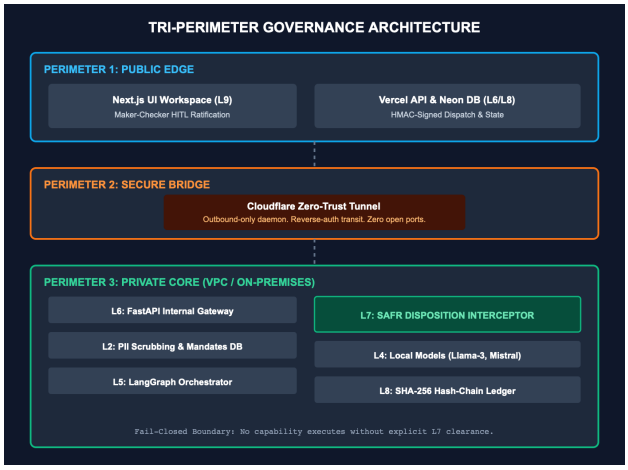


Figure 1: The Tri-Perimeter Governance Architecture mapping Edge, Bridge, and Private Core layers. © Balian AI

At the heart of the Private Core lies the principle of *Continuous Trajectory Governance*. The agentic orchestrator (Lang Graph, L5) manages task trajectories but has zero authority to execute tools directly. Every proposed invocation must be routed down to the SAFR Disposition Interceptor (L7) for deterministic evaluation before execution.

4. Operationalising SAFR Guidance for Social Care

Under the derived guidance from the Monetary Authority of Singapore's (MAS) Safeguards for Agentic Finance at Runtime (SAFR) framework, this architecture translates financial-grade runtime safeguards into human services. SAFR was introduced as an industry reference approach—not a monolithic algorithm or formal regulation—to define

Copyright © Shasana Balian AI Pte Ltd 2026

how agent actions should be verified at the point of decision. In social services, this runtime governance checkpoint consists of four interconnected components: Agent Identity, Controls Repository, Disposition Engine, and Hash-Chain Audit Log.

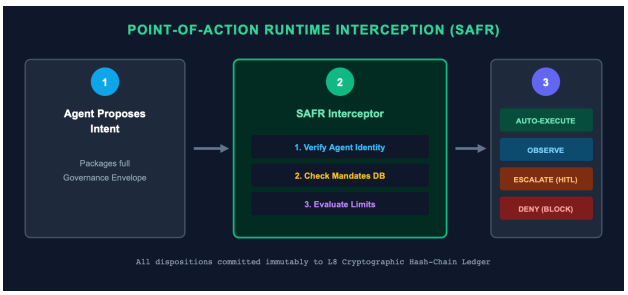


Figure 2: SAFR point-of-action runtime interception flow bridging proposed intents to safe execution.

5. Real-World Use Cases in Human Services

Domestic and Family Violence (DFV) & Crisis Support: When generating safety plans or processing intake triage for DFV victims, the Enterprise Sanitisation layer (L2) conducts zero-trace PII scrubbing within a private core, protecting individuals from digital surveillance or tracking by abusive partners.

- **Disability & Assistive Tech (AAC):** For neurodivergent individuals using AI as an adaptive communication tool, local-first architecture provides low-latency execution without cloud-induced lag and protects users from commercial behavioural profiling.

Public Sector Casework & Child Protection: Statutory child protection data requires absolute sovereignty. Under SAFR, agents cannot unilaterally authorise interventions. Breaches trigger a Tiered Containment Freeze, escalating the case to human maker-checkers alongside SHAP/LIME Explainability metrics.

6. Advancing Defensible, Humane AI

The transition from pilot phase curiosity to enterprise-grade social care requires operational discipline and defensible governance. Balian AI stands at the forefront of this strategic pivot. Co-founded by experts in high-stakes regulated banking transformations and enterprise delivery at scale, Balian AI embodies the principle of "Applied AI, governed from strategy to operations." By designing and running private core architectures equipped with continuous runtime SAFR interception, the design ensures that every AI action is bounded by a written mandate, approved by a person where it matters, and kept strictly on the record.

References & Citations

1. MAS. Safeguards for Agentic Finance at Runtime (SAFR), BuildFin.ai, 2026.
2. MAS. Guidelines on Artificial Intelligence Risk Management, Oct 2026.
3. IMDA. Model AI Governance Framework for Agentic AI, 2026.
4. Balian AI. Company Profile: Applied AI, governed from strategy to operations, 2026.
5. Perplexity Hub. A Local-First Agent for Knowledge Work, 2026.
6. NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.
7. Balian AI: https://balian.ai

Document dated: 14th October 2026