Decision-OS V5 SiriusA: Zero-Knowledge Confirmation Layer for the Protection of Life

Shinichi Nagata
Kanagawa, Japan
siriusa.paper@gmail.com

SSOT: GitHub decision-os-paper, commit840c85de344f5dd197dc5122ad9f4ba452f2970d.

This PDF corresponds exactly to this commit.

Author Responsibility Declaration

The author assumes sole responsibility for the research design, claims, evaluation metrics, operating point selection (Quiet/Standard/Aggressive), and the definition of the 10-second ritual presented in this work, as well as for all publication decisions.

AI systems (OpenAI GPT-5 Thinking as the primary model and Google Gemini Ultra DeepResearch as the secondary) were used to generate alternatives and refine language. All acceptance, rejection, and editing were performed exclusively by the human author, with no modification of numerical data or methods.

This study remains within a non-medical scope, adheres to the principles of no automatic transmission, payment, or reporting, and preserves human control through two-step confirmation and a revoke code.

We add a Zero-Knowledge family confirmation layer that verifies authorization (m-of-k) without PII, couples duress-aware thresholds (θ1, θ2, θ_critical) with Δt control, and preserves two-step/revoke with non-PII evidence sealing (ZIP + SHA-256).

The North Star of the research is expected harm minimization. Reproducibility and accountability are ensured through evidence ZIP + SHA256, non-PII KPIs, weekly SPC (statistical process control), and calibration (Brier/ECE).

All artifacts are anchored to SSOT commit `840c85de344f5dd197dc5122ad9f4ba452f2970d`.

1. Introduction

SiriusA Core (Central Claim)

This work proposes an operational OS that places the protection of life as its highest objective and aims to minimize fatal delay.

It formalizes 10-second behavioral support as a human–AI shared decision-making process.

Contributions

1. A time-optimization equation Teff=Δtset+Pres−1T_\mathrm{eff} = \Delta t_\mathrm{set} + P_\mathrm{res}^{-1}Teff=Δtset+Pres−1 with domain-specific Δt.
2. Integration of family multisig (k/n, Δt).
3. Implementation standards E-1 to E-8.
4. KPI framework and audit design.

Definitions (this section only)

- **SiriusA:** An aspirational design architecture — a design layer dedicated to preserving human dignity and autonomy.
- **North Star:** The objective function — an indicator that maximizes sustained benefit without compromising human decision authority.

Objective

Guided by protection of life as the highest aim, the system seeks to minimize fatal delay while maintaining human control through the 10-second ritual and two-step confirmation.

1. Define effective time (Teff=Δtset+Pres−1T_\mathrm{eff} = \Delta t_\mathrm{set} + P_\mathrm{res}^{-1}Teff=Δtset+Pres−1) to decompose and optimize sources of delay.

2. Integrate approval and time windows via family multisig (k/n, Δt) to standardize safe operations.
3. Audit and calibrate performance using non-PII KPIs, evidence ZIP + SHA256, weekly SPC (statistical process control), and calibration (Brier/ECE).

a) Formula: Teff=Δtset+Pres−1T_\mathrm{eff} = \Delta t_\mathrm{set} + P_\mathrm{res}^{-1}Teff=Δtset+Pres−1; Figure 1: family multisig BPMN/UML.

b) KPIs: Adherence, Decision Time, FPR, FNR, Net Benefit, Brier, ECE.

c) Logs: two-step confirmation, revoke code, signer ID. Direct-call phrases (119/110) are exceptions to the WAIT48h rule, but two-step confirmation remains required.

Pitfalls (≤2)

- Automatic transmission, payment, and reporting are excluded to ensure clear responsibility boundaries and suppress false alarms.
- The operating point (Quiet/Standard/Aggressive) is chosen by the human; any automatic threshold proposal requires explicit human approval.

This paper positions protection of life as its North Star and implements human–AI shared decision-making as 10-second behavioral units in safety-critical, non-medical domains such as fraud prevention, disaster response, fall detection, and family communication.

To reduce fatal delay, effective time Teff=Δtset+Pres−1T_\mathrm{eff} = \Delta t_\mathrm{set} + P_\mathrm{res}^{-1}Teff=Δtset+Pres−1 is defined, treating both the approval window (Δt) and responsiveness (P_res) as design variables.

The architecture centers on human final consent, disallowing any automatic transmission, payment, or reporting that bypasses the two-step confirmation and revoke code.

Direct-call phrases (119/110) are exempt from the WAIT48h rule, but two-step confirmation remains mandatory.

Family multisig (k/n, Δt) defines an approval structure—including a single-member 1/1 fallback—and delegates the choice of operating point on the ROC curve to each family's value function.

Evaluation employs non-PII KPIs (Adherence, Decision Time, FPR/FNR, Net Benefit, Brier/ECE) and publishes tamper-resistant evidence via evidence ZIP +

SHA256.

Weekly SPC monitors drift, while decision curve analysis and calibration metrics validate practical effectiveness.

Together, SiriusA (aspirational design) and the North Star (objective function) form a unified operational framework that advances expected harm minimization.

Guide: See §4 Architecture Overview for the overall structure and §2 Problem Statement and Scope for definitions and operational boundaries.

2. Problem Statement and Scope (Non-medical)

SiriusA Core — Scope and Claims

This work is limited to safety-critical, non-medical domains: fraud prevention, disaster response, fall events, and family communication.

AI does not replace human judgment. Human approval is preserved via **two-step confirmation** and a **revoke code**.

We define **fatal delay** as the total time from hazard detection to action completion, and make explicit the design boundaries that minimize it.

Objective

In non-medical settings where AI does not act autonomously, clarify the boundary of human–AI collaboration and the structure of fatal delay.

Fixed operational boundaries

- Scope: non-medical. Default **operating point (Quiet/Standard/Aggressive)** is **Quiet**.
- **No automatic transmission, payment, or reporting.** All external actions must pass **two-step confirmation** and remain reversible with a **revoke code**.
- **Direct-call phrases (119/110)** are exceptions to **WAIT48h**, but **two-step confirmation** remains required. See **E-1 to E-8 (§3)** for implementation details.

1. Constraining the scope to non-medical domains reduces legal/ethical risk and improves design reproducibility.

2. A hold-until-human-approval structure balances human responsibility with AI **responsiveness**.
3. Decomposing **fatal delay** into time components surfaces mathematical improvement points on both the system and human sides.

a) **Time model (Equation 1 reused):** (T_\mathrm{eff} = \Delta t_\mathrm{set} + P_\mathrm{res}^{-1})

b) **KPIs:** Decision Time, FPR, FNR, Net Benefit (expected harm minimization metric)

c) **Logs:** approval timestamps, revoke code entries, and approval rates for **family multisig** (k/n, Δt)

Pitfalls (≤2)

- Ambiguous scope expands AI's intervention range and erodes human responsibility boundaries.
- Over-prioritizing responsiveness invites designs that skip human approval and breach the ethical boundaries (E-1 to E-8).

The target domains directly affect life yet are not medical practice. AI's role is assistance—not substitution—under the premise that humans remain the final actors. The system accelerates human decision-making to reduce **fatal delay**, defined as the elapsed time from hazard to completed action, consisting of: (i) AI detection/presentation delay, (ii) human perception/approval delay, and (iii) execution delay. What AI can optimize ends at the presentation component within (i) and (ii); final approval and action remain human responsibilities.

SiriusA renders this boundary explicit through **two-step confirmation** and a **revoke code**, while disallowing automatic transmission, payment, and reporting. **Family multisig (k/n, Δt)** adapts approval thresholds and **approval windows (Δt)** to household structure, supporting human approval during emergencies. This preserves AI **responsiveness (P_res)** while allowing humans to choose the balance between **false positive rate (FPR)** and **false negative rate (FNR)**. This scope definition is foundational; the following sections build on the time-decomposition model to present the ethical design principles **E-1 to E-8** and the selection of the **operating point (Quiet/Standard/Aggressive)**. *ZK is used solely for qualification proofs (authorization without PII); the system never auto-initiates transfers, payments, or reports.*

Guide: See §4 for the architectural overview.

---

3. Design Principles and Ethical Boundaries (E-1 to E-8)

Guide: This chapter defines the principles (E-1 to E-8) only. For metrics and methods, see **§9**.

SiriusA Core (Pillars)

We define eight ethics (E-1 to E-8), centered on human final approval, as the design principles for AI operations.

By constraining automation and institutionalizing **two-step confirmation** and a **revoke code**, human choice is preserved.

We adopt **non-PII KPIs** and **evidence ZIP + SHA256** so that adherence to ethical boundaries is implemented in an auditable way.

Objective

Make explicit the ethical boundaries (E-1 to E-8) in the SiriusA architecture and specify the limits and assurance conditions for AI operations.

1. Use **two-step confirmation** and a **revoke code** to prevent misfires and mis-approvals, preserving human ultimate responsibility.
2. Allow humans to choose the **operating point (Quiet/Standard/Aggressive)** to sustain ethical autonomy.
3. Rely on non-PII data for KPI operations and publish **evidence ZIP + SHA256** to guarantee accountability and transparency.

a) **Figure:** E-1–E-8 framework (ethical boundaries and responsibility demarcation)

b) **Metrics/Analysis:** See **§9** (KPI definitions, calibration, decision curves, SPC)

c) **Logs:** revoke code history, approval timestamps, ZIP + SHA256 signature list

Pitfalls (≤2)

- Relaxing any one of E-1–E-8 can introduce ethical failure points even under Quiet mode.
- Without transparency, AI decisions become unauditable and social trust is eroded.

SiriusA pursues **protection of life** while minimizing AI automation by defining eight ethics (E-1–E-8) as its central principles.

E-1–E-4 cover **human-centered design**, while E-5–E-8 address **audit and transparency**.

- **E-1:** two-step confirmation
- **E-2:** revoke code
- **E-3:** human-selected modes (Quiet/Standard/Aggressive)
- **E-4:** 119/110 WAIT48h exception rule
- **E-5:** non-PII KPIs
- **E-6:** evidence ZIP + SHA256
- **E-7:** weekly SPC (statistical process control) for ethical-drift monitoring
- **E-8:** auditable logs and disclosure responsibilities

By systematizing these, all AI behavior is defined as **collaboration contingent on human consent**.

The architecture enforces **constraints by structure, not by prohibition**. For example, a two-stage **revoke code** implementation provides structural braking so that no AI output is transmitted externally without user approval.

The three modes—Quiet, Standard, and Aggressive—are **operating points** chosen by humans under **expected harm minimization**, allowing ethical management of the trade-off between **responsiveness (P_res)** and safety.

Operations are evaluated using **non-PII KPIs**, and transparency for third-party audit is ensured by publishing **evidence ZIP + SHA256**.

Thus, the framework functions as an **implementation rule set (E-1–E-8)** with auditability and reproducibility—not merely an ethical declaration.

It preserves a practical balance that minimizes **fatal delay** while keeping within ethical boundaries throughout human–AI collaboration.

4. Architecture Overview (SiriusA)

Roadmap: This chapter presents the **SiriusA overview**. Details appear in **§5 (time optimization: T_eff, P_res)**, **§6 (HRI/UX: the 10-second ritual and the**

five-line UI), and **§7 (family multisig, k/n, Δt)**. **Design principles are in §3**, **operational audit in §8**, and **evaluation/metrics in §9**.

SiriusA Core (Pillars)

- A serial pipeline centered on the **10-second ritual**: **Trigger Detection → Mode Management → Five-line UI → Family one-liner → Two-step confirmation → Signed logging**.
- Humans choose the **operating point (Quiet/Standard/Aggressive)** to jointly minimize TeffT_\mathrm{eff}Teff and suppress false alarms.
- **Evidence ZIP + SHA256** and **signer-ID logs** make operations auditable via **non-PII KPIs**.

Objective

Provide a processing pipeline and responsibility demarcation that minimize **fatal delay** without compromising human-led decision-making.

1. The serial pipeline (Detect → Mode → UI → Family one-liner → Confirm → Log) enables decomposition and optimization of TeffT_\mathrm{eff}Teff.
2. The **operating point (Quiet/Standard/Aggressive)** is selected by humans as a point on the **receiver operating characteristic (ROC) curve**, adapted by tuning PresP_\mathrm{res}Pres and Δt\Delta tΔt.
3. **Auditability** (non-PII KPIs / evidence ZIP + SHA256 / signer IDs) functions as an implementation rule set that upholds the ethical boundaries **E-1–E-8**.

a) **Figure 1:** BPMN/UML (SiriusA loop around the perimeter). **Equation:** Teff=Δtset+Pres−1T_\mathrm{eff} = \Delta t_\mathrm{set} + P_\mathrm{res}^{-1}Teff=Δtset+Pres−1

b) **KPIs:** Decision Time, Adherence, FPR, FNR, Net Benefit, Brier, ECE

c) **Logs:** two-step confirmation, revoke code, timestamps for **direct-call phrases (119/110)** (WAIT48h exception), and signer IDs

Pitfalls (≤2)

- Permitting automatic transmission, payment, or reporting risks bypassing human approval and collapsing responsibility under false alarms.
- Exceeding five UI lines degrades PresP_\mathrm{res}Pres and worsens TeffT_\mathrm{eff}Teff.

SiriusA implements a **serial pipeline** with the **10-second ritual** at its core:

1. **Trigger Detection:** Lightweight detection across fraud, disaster, falls, and family communication; over-detection is absorbed by the **Quiet** default.
2. **Mode Management:** Humans select **Quiet/Standard/Aggressive** to align the **FPR–FNR** trade-off with the family's value function.
3. **Ritual UI (≤5 lines):** Present the required action in five lines or fewer, plus a **family one-liner** to manage cognitive load.
4. **Two-step Confirmation:** All transfers, outreach, and reporting pass **two-step confirmation**; a **revoke code** reverses the most recent approval. **Direct-call phrases (119/110)** are exempt from **WAIT48h**, but **two-step** still applies.
5. **Evidence & Logging:** Capture **non-PII KPIs** and ensure auditability with **evidence ZIP + SHA256** and **signer IDs**.

This architecture decomposes and implements Teff=Δtset+Pres−1T_\mathrm{eff} = \Delta t_\mathrm{set} + P_\mathrm{res}^{-1}Teff=Δtset+Pres−1.

The **approval window (Δt_set)** is governed by **two-step confirmation** and **family multisig (k/n, Δt)**; **responsiveness (P_res)** is raised by the five-line UI, **direct-call phrases (119/110)**, the family one-liner, and routinized procedures.

A **Quiet** default prevents externalization of false alarms; **Standard/Aggressive** are selected explicitly when time is more valuable. **Figure 1 (BPMN/UML)** depicts the flow, with an outer **weekly SPC** feedback loop to detect drift and propose threshold updates (subject to human approval).

In this way, 10-second **shared decision-making** is made practicable without compromising human final consent, responsibility boundaries, or auditability.

For gate placement and branching in deployment, **see §7, Figure 1**.

5. Time-Optimization Model and Operating-Point Selection

Table 1. Domain-specific Δt Presets

Domain	Default operating point	Recommended Δt (sec)	Notes
Fraud	Quiet	90–180	High false-approval cost; few notifications, multiple two-step confirmations.
Disaster	Standard	15–45	Prioritizes immediate response; shortened message templates.
Fall	Quiet	120–240	Extend Δt at night to suppress false alarms.
Family communication	Quiet	Flexible	Adaptive based on time-of-day presets.

Guide: This chapter details the **time optimization** component of §4. For boundaries, see **§2**; for metrics and analysis, see **§9**.

SiriusA Core (Pillars)

At the center is **effective time (T_eff)**, defined as:

Teff=Δtset+Pres−1T_\mathrm{eff} = \Delta t_\mathrm{set} + P_\mathrm{res}^{-1}Teff=Δtset+Pres−1

This formulation minimizes **fatal delay** along two axes: the **approval window (Δt)** and **responsiveness (P_res)**.

Humans select among **Quiet, Standard, and Aggressive** operating points—treated as positions on the **receiver operating characteristic (ROC) curve**—based on their value function.

Each domain (fraud, disaster, fall, family communication) has tailored Δt presets and default modes to achieve **expected harm minimization**.

Objective

To present a mathematical time structure and an optimal **operating-point model** that minimizes **TeffT_\mathrm{eff}Teff** while preserving human judgment.

1. TeffT_\mathrm{eff}Teff combines a **design variable** (Δt) and a **behavioral variable** (P_res), both controllable in design and operation.
2. Humans select the **operating point (Quiet/Standard/Aggressive)** on the ROC curve; the **Quiet** default suppresses **false positives (FPR)**.
3. Domain-specific Δt presets balance **FPR/FNR** trade-offs according to time value.

Implementation flow and gate layout: **see §7, Figure 1 (family multisig).**

a) **Formula (Equation 1 restated):** Teff=Δtset+Pres−1T_\mathrm{eff} = \Delta t_\mathrm{set} + P_\mathrm{res}^{-1}Teff=Δtset+Pres−1

Figure: Operating points (Quiet/Standard/Aggressive) on the ROC curve.

b) **KPIs:** Decision Time, FPR, FNR, Net Benefit, Brier, ECE.

c) **Logs:** approval timestamps, Δt exceedance rates, two-step/revoke code/family (k/n, Δt) execution records.

Pitfalls (≤2)

- Setting Δt too short → approval incomplete (FNR ↑); too long → worsens TeffT_\mathrm{eff}Teff.
- Excessive UI complexity or notifications **degrade PresP_\mathrm{res}Pres** and undermine Quiet-mode stability.

SiriusA defines **effective time (T_eff)** to minimize **fatal delay**, where the **approval window (Δt_set)** is a design constant and **responsiveness (P_res)** a variable shaped by human–UI interaction.

Δt is managed through **two-step confirmation** and **family multisig (k/n, Δt)**, while **P_res** is increased by the **five-line UI**, the **family one-liner**, redundant notification paths, and **direct-call phrases (119/110)** (WAIT48h exceptions).

The **operating point (Quiet/Standard/Aggressive)** represents a human-selected position on the ROC curve:

- **Quiet:** prevents externalization of over-detections.
- **Standard:** used when time value is high.
- **Aggressive:** reserved for imminent life-threatening situations.

AI never auto-selects the mode; humans switch manually, with reversibility ensured by the **revoke code**.

Each domain's temporal characteristics yield a distinct optimal Δt — fraud (moderate/Quiet), disaster (short/Standard), fall (long/Quiet), family communication (flexible).

This differentiation enables domain-specific **expected harm minimization**.

Weekly SPC tracks drift in **TeffT_\mathrm{eff}Teff** using **non-PII KPIs** (Decision Time, FPR, FNR, Brier/ECE), producing human-reviewed adjustment proposals for Δt and PresP_\mathrm{res}Pres. Each approval is archived as **evidence ZIP + SHA256**. AI may propose, but humans decide.

This **human-chosen optimum** makes time optimization practical while preserving the ethical boundaries (E-1–E-8).

It forms the core of the **SiriusA time optimization model**.

6. Interaction Design (HRI/HF)

Guide: This chapter details **HRI/UX** within §4. See **§2** for boundaries and **§9** for metrics and methods.

SiriusA Core (Pillars)

We fix **shared decision-making** with humans as the final decision-makers through the **10-second ritual**. AI is constrained to **presentation, visualization, and recording**, with the aim of minimizing **fatal delay** in service of **protection of life**.

Objective

Deliver a UI/UX that increases **PresP_\mathrm{res}Pres** and reduces **TeffT_\mathrm{eff}Teff**, while leaving **FPR/FNR** under human control.

1. **Five-line UI — [Input → Recheck → Approve → Revoke → Still] —** implements the **10-second ritual**, with the **revoke code** always available.
2. **Two-step confirmation — 1st input → 2-second cool-down → 2nd commit**. The assist layer shows the remaining **approval window (Δt)** and the current **operating point (Quiet/Standard/Aggressive)**. See **§7, Figure 1**.
3. **Family multisig (k/n, Δt)** with time-of-day presets optimizes **Δt**, notification volume, and input modalities (audio/light/haptic).

a) **KPIs:** Decision Time, **PresP_\mathrm{res}Pres**, FPR, FNR, Net Benefit, calibration (Brier/ECE)

b) **Analysis:** decision curve analysis; weekly SPC for drift detection

c) **Logs:** non-PII KPIs with **evidence ZIP + SHA256** (two-step timestamps, revoke usage, family **k/n** attainment rate)

Pitfalls (≤2)

- Too much guidance → cognitive load ↑ → **PresP_\mathrm{res}Pres** ↓ (do not exceed five lines).
- Fixed presets → situational mismatch (weekly threshold proposals → apply only after human approval).

We implement UI/UX that simultaneously optimizes **TeffT_\mathrm{eff}Teff** and **PresP_\mathrm{res}Pres** without compromising human final approval. The **10-second ritual** is central. The **five-line UI** comprises **[Input → Recheck → Approve → Revoke → Still]**, with the **revoke code** available at every stage. **Two-step confirmation** inserts a 2-second recognition cool-down after the first input and completes on the second commit. The assist layer continuously displays the remaining seconds in the **approval window (Δt)** and the **operating point (Quiet/Standard/Aggressive)**; **Quiet** is the default to avoid externalizing over-detections, while **Standard/Aggressive** are human-selected as context demands.

Treat **family multisig** as an action token with **k/n** and **Δt**. Default to in-home **2-of-2** (phone + watch/stationary). Time-slot presets (late night / early morning / daytime / evening / pre-sleep) adjust **Δt**, notification intervals, and cue modalities. Extend **Δt** at night; shorten intervals during daytime—empirically raising **PresP_\mathrm{res}Pres**. **Direct-call phrases (119/110)** are exceptions to **WAIT48h**, but **two-step** remains; the system never performs automatic transmission, payment, or reporting.

All observation uses **non-PII KPIs**. We preserve **Decision Time, Adherence, revoke activation rate, FPR, FNR** in **evidence ZIP + SHA256**. **Weekly SPC** monitors drift; **calibration (Brier/ECE)** aligns reliability; and **decision curve analysis** quantifies net benefit across operating points. Updates follow **proposal → human approval**, maintaining reversibility. In this way, humans can raise **PresP_\mathrm{res}Pres** while shortening **TeffT_\mathrm{eff}Teff**, retaining ongoing control over **operating point, Δt, and k/n** to minimize **fatal delay**.

---

7. Family Multisig and Figure 1 (BPMN/UML)

Guide: Figure 1 is primarily referenced in this chapter. Other sections should cite **"§7, Figure 1."**

SiriusA Core (Pillars)

We design **family multisig** as a **cascading hierarchy** that curbs **fatal delay** while minimizing erroneous intervention.

Each Tier specifies an **approval window (Δt)**, **retry count (r)**, and **consecutive-miss threshold (k)**, with the **operating point** chosen by humans.

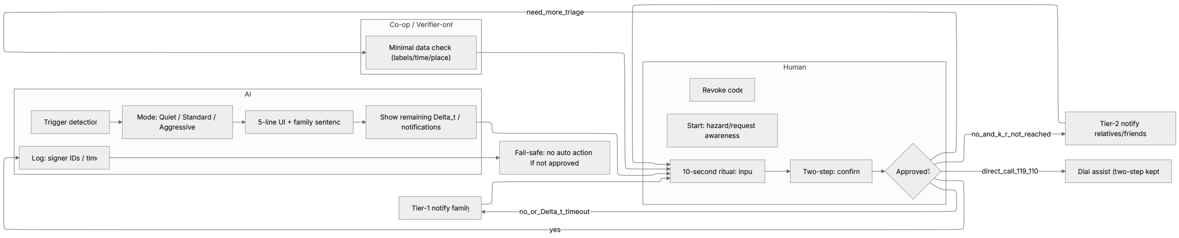
need_more_triage
Co-op / Verifier-onl
Minimal data check (labels/time/place)
AI
Trigger detectior
Mode: Quiet / Standard / Aggressive
5-line UI + family sentenc
Show remaining Delta_t / notifications
Log: signer IDs / tim
Fail-safe: no auto action if not approved
Tier-1 notify family
no_or_Delta_t_timeout
yes
Human
Revoke code
Start: hazard/request awareness
10-second ritual: inpu
Two-step: confirn
Approved?
no_and_k_r_not_reached
Tier-2 notify relatives/friends
direct_call_119_110
Dial assist (two-step kept

Figure 1. family multisig (BPMN/UML-style).

Three lanes (Human / AI / Co-op) with cascading flow (Tier-0 → Tier-4).

The legend lists **Δt, r, k, revoke**. **Unapproved ⇒ no automatic action.**

Direct-call phrases (119/110) are **WAIT48h exceptions**, yet **two-step confirmation** still applies.

Objective

Connect in-family consensus to the **10-second ritual**, defining a safe approval chain that raises **PresP_\mathrm{res}Pres** while reducing **TeffT_\mathrm{eff}Teff**.

1. **Hierarchy:** Tier-0 self → Tier-1 family → Tier-2 relatives/friends → Tier-3 Co-op (Verifier-only) → Tier-4 direct-call (119/110).
2. **1/1 & 1/2 safeguards:** default **in-home 2-of-2** (phone + watch/stationary). At night, increase **Δt**, widen notification intervals, and reallocate **k/r**.
3. **Co-op:** activates **only on trigger**, validates **minimal data** (label/time/place), and has **no reporting authority**, preserving ethical boundaries.

a) **Figure 1:** BPMN/UML swimlanes (Human / AI / Co-op) with cascading branches (**Δt, r, k**).

b) **KPIs: PresP_\mathrm{res}Pres**, FNR, FPR, cross-Tier penetration failure rate, Decision Time (weekly SPC; calibration = Brier/ECE).

c) **Logs:** non-PII KPIs + **evidence ZIP + SHA256** (approval timestamps, **revocation latency**, k/n attainment rate).

Pitfalls (≤2)

- Excessive Tiers inflate **TeffT_\mathrm{eff}Teff** (keep **Quiet** as default to mitigate).
- Making Co-op monitoring continuous undermines trust (restrict to **on-trigger verification**).

We implement **family multisig** as a cascading mechanism. Beginning with the **Tier-0 10-second ritual**, escalation proceeds **only** when approval is at risk of exceeding the **approval window (Δt)**.

Tier-1 (family) increases **PresP_\mathrm{res}Pres** with familiar routines and a family one-liner; only when **k** consecutive misses and **r** retries fail do we fallback to **Tier-2 (relatives/friends)**.

Tier-3 (Co-op; Verifier-only) is **triggered on demand**, validating **label/time/place** as minimal evidence to filter false alarms, while holding **no authority to report**.

Tier-4 (direct-call phrases 119/110) is permitted as a **WAIT48h exception**; **two-step confirmation** and the **revoke code** remain in place.

To mitigate **1/1** and **1/2** risks, we default to **device-level 2-of-2** (phone + watch/stationary).

Night-time presets extend **Δt**, lengthen notification intervals, and retune **k/r**.

Quiet is the default; **Standard/Aggressive** are explicitly human-selected **operating points**.

We monitor with **non-PII KPIs** (Decision Time, **PresP_\mathrm{res}Pres**, FPR, FNR, cross-Tier failure rate) via **weekly SPC** and **calibration (Brier/ECE)**.

Threshold updates are **proposed by AI**, **approved by humans**, and archived via **evidence ZIP + SHA256**.

Figure 1 (BPMN/UML) places **Δt, r, k, revoke** across the Human/AI/Co-op lanes, visualizes cascading branch conditions, and enforces the fail-safe **“unapproved = no automatic action.”**

In doing so, we minimize **fatal delay** while preserving **shared decision-making**.

Note: Explicitly include **Δt, r, k, revoke** in the Figure 1 legend during figure creation.

§7.3 Zero-Knowledge Family Confirmation Layer (ZK 310/320)

Overview. We introduce a Zero-Knowledge (ZK) family confirmation layer to verify *who is authorized to approve* without exposing any personally identifiable information (PII). The layer augments the existing two-step confirmation, revoke code, and family multisig (m-of-k) with a cryptographic qualification check. Only the proof that "an authorized subset approved" is revealed; no keys, identities, or raw credentials are disclosed. *(See Fig. 1 (ZK 310/320).)*

Design. The layer consists of (i) a ZK authentication tier (310) that issues a boolean flag `zk_qual` indicating that an m-of-k threshold was satisfied by authorized parties, and (ii) a ZK signature-verification tier (320) that attests each approval was produced by a policy-compliant signing context. The system logs a non-PII audit tuple ⟨time, event-type, `zk_qual`, Δt, decision-state⟩, sealed in an evidence ZIP with SHA-256.

Duress-aware control. Let `duress_score ∈ [0, 1]` and thresholds $\theta_1 < \theta_2 < \theta\_critical$. When `duress_score ≥ θ_critical` and `zk_qual = true`, the controller sets Δt → 0 for immediate stop/hold (fatal-decision prevention). For $\theta_2 \leq$ `duress_score` < θ_critical, the controller expands the approval window and demands higher m-of-k (e.g., from 2-of-3 to 3-of-4) before execution. For `duress_score` $< \theta_1$, the baseline operating point applies.

Safety invariants. (S1) Two-step confirmation is never bypassed. (S2) The revoke code remains valid throughout the approval window. (S3) Emergency channels (e.g., medical or police) are exempt from WAIT/flip delays yet still pass through the confirmation view. (S4) The ZK layer never initiates automatic transfers or payments.

Privacy & audit. The ZK layer records outcome flags only. Identity-bearing material and raw signatures never leave the user domain. All artifacts are bundled as an evidence ZIP (metadata-only) and hashed (SHA-256) for integrity and reproducibility.

Interoperability. The layer is protocol-agnostic: any wallet, device, or verifier that can emit or verify standard ZK proofs and policy-scoped signatures can participate. BPMN flow aligns with the existing approve → revoke → execute → hold pipeline; state transitions follow the same Δt semantics.

Effect. Under coercion or social engineering, users often “consent” themselves into loss. By separating **qualification** (who may approve) from **content** (what is revealed), the ZK layer preserves autonomy while minimizing false positives and negatives on critical halts.

8. Safety Verification and Operational Audit

Guide: This chapter focuses strictly on **operational audit** (SPC and falsification patches). For metrics and analytic methods, see **§9**.

SiriusA Core (Pillars)

Safety is not a pre-release event. Within **shared decision-making**, we continuously reduce **fatal delay** by centering **non-PII KPIs** and **evidence ZIP + SHA256**, and by running **weekly** monitoring and corrective actions.

Objective

Collect **non-PII KPIs**, package them into a ZIP, publish the hash, and use **weekly SPC** plus **falsification patches** to adjust the **operating point (Quiet/Standard/Aggressive)** and the **approval window (Δt) — only with human approval**.

1. **Observe:** Record the **minimal set — T_eff, P_res, FPR, FNR, False Alarm Burden (person-hours)** — and seal settings/logs in **evidence ZIP + SHA256**.
2. **Monitor:** Use weekly statistical process control to detect drift; run **calibration (Brier/ECE)** and net-benefit comparisons **as defined in §9**.
3. **Correct:** Apply **minimal-width patches** (thresholds, UI wording, notification counts) via **propose → human two-step confirmation → apply**. The AI never changes settings on its own.

a) **Observed items:** T_eff; P_res; FPR; FNR; False Alarm Burden (person-hours)

b) **Metrics & analysis: See §9** (KPI definitions, calibration, decision curves, SPC)

c) **Traceability:** two-step timestamps; **revocation latency**; family multisig **k/n** attainment — archived as **non-PII KPIs + ZIP + SHA256**

Pitfalls (≤2)

- Tracking too many indicators increases cognitive and privacy burden — **stick to the minimal set**.
- Turning monitoring into “always-on” violates ethical boundaries — **Co-op is Verifier-only and trigger-only**.

We define safety as an operational loop — **measure, compare, correct**. Observations are limited to **non-PII KPIs**, capturing **effective time (T_eff)**, **responsiveness (P_res)**, **false negative rate (FNR)**, and **False Alarm Burden (person-hours)**. Configuration and logs are sealed with **evidence ZIP + SHA256** to ensure tamper resistance. We conduct **weekly SPC** to detect drifts such as extended delays, excessive revocations, or declining responsiveness. Analytical steps — calibration, net-benefit comparisons, and threshold evaluation — follow **§9**.

Corrections are applied as **falsification patches**, making **minimal** updates to the **approval window (Δt)**, notification counts, and UI text/order. Every change requires **human approval** using **two-step confirmation** and a **revoke code**. The system never performs **automatic transmission, payment, or reporting**. **Direct-call phrases (119/110)** are **WAIT48h** exceptions, but **two-step confirmation** remains.

For **family multisig**, we log **k/n** and **Δt** performance (attainment rate, Δt-exceedance rate, **revocation latency**). Threshold proposals follow **propose → approve → apply**, lowering **fatal delay** while preserving **protection of life** and enabling deviations, degradation, and false events to be tracked with **reproducible logs**.

— **8.1 Collection → ZIP → hash (non-PII KPIs).** Gather the minimal KPI set and preserve as **evidence ZIP + SHA256** (definitions/calculations per **§9**).

— **8.2 Weekly monitoring & falsification patches.** Detect drift with **weekly SPC**; apply patches **only after human approval** (**no automatic changes**).

— **8.3 Auditability (recomputation).** Align two-step, revoke, **k/n**, **Δt**, and **operating point (Quiet/Standard/Aggressive)** so that third parties can **recompute and match** (procedures in **§9**).

---

9. Evaluation Plan (KPIs/Protocols)

Canon: KPI names, definitions, acquisition procedures, and analytical methods (**calibration, decision curve analysis, SPC**) are defined **exclusively** in this chapter.

KPIs

1. **Adherence:** Percentage of steps executed exactly as specified.
2. **Decision Time:** Seconds to complete approval.
3. **PresP_\mathrm{res}Pres** (responsiveness): Latency from presentation to user action (higher implies faster response).
4. **FPR (false positive rate):** Proportion of false positives.
5. **FNR (false negative rate):** Proportion of false negatives.
6. **False Alarm Burden (person-hours):** Total human time spent handling false alarms.
7. **Net Benefit (decision curve):** Net benefit as a function of threshold probability.
8. **Calibration (Brier/ECE):** Consistency between predicted and observed outcomes.

SiriusA Core (Pillars)

We design evaluation as **case × operating point (Quiet/Standard/Aggressive) × time-of-day** and use reproducible comparisons—based on **non-PII KPIs** and **evidence ZIP + SHA256**—to jointly assess **fatal delay**, **FPR/FNR**, and **Net Benefit** in service of **protection of life**.

Objective

Validate case-wise **expected harm minimization** and propose recommended ranges for the **operating point (Quiet/Standard/Aggressive)** and the **approval window (Δt)**.

1. **Case design:** Four domains—fraud, disaster, fall, family communication—three trials each (12 total). Time-of-day balanced across **[late night / daytime]**; order effects controlled via a **Latin square**.
2. **KPI acquisition:** Adherence, Decision Time (sec), PresP_\mathrm{res}Pres, FPR, FNR, False Alarm Burden (person-hours), Net Benefit, **calibration (Brier/ECE)**.

3. **Analysis: Decision curve analysis** compares net benefit across operating points; **weekly SPC** monitors drift. Threshold proposals are updated only with **human approval**.

a) **Design:** Factorial plan (case × operating point × time-of-day); pre-register **Δt, k/n, notification counts**.

b) **Metrics:** Adherence; Decision Time; FPR; FNR; PresP_\mathrm{res}Pres; Net Benefit; Brier; ECE.

c) **Evidence:** Log non-PII KPIs and settings as **evidence ZIP + SHA256** (with pre-registration ID and commit SHA).

Pitfalls (≤2)

- Fixed sequence Quiet → Standard → Aggressive induces learning bias (**use a Latin square**).
- Not measuring **revocation latency** inflates **PresP_\mathrm{res}Pres** (**record continuously**).

We evaluate 12 tasks: four domains × three trials. Each trial assigns **Quiet/Standard/Aggressive** using a **Latin square** to control order effects. Time-of-day is balanced between **late night** and **daytime**. **Approval window (Δt)**, **family multisig (k/n, Δt)**, and **notification counts** are **pre-registered** and fixed. Inputs are limited to the **10-second ritual (five-line UI)**; every run must pass **two-step confirmation**, with a **revoke code** available.

Collected **non-PII KPIs** include **Adherence**, **Decision Time** (seconds to approval), **PresP_\mathrm{res}Pres**, **FPR**, **FNR**, **False Alarm Burden (person-hours)**, **Net Benefit**, and **calibration (Brier/ECE)**. **Direct-call phrases (119/110)** remain **WAIT48h exceptions**, yet **two-step** is maintained; **no automatic transmission, payment, or reporting** is ever used. Missing approvals are treated as **Δt exceedances**, and **unapproved = no automatic action**. **Revocation latency** is defined as the seconds from **revoke code** entry to completion of invalidation.

Analytically, **decision curve analysis** compares **Net Benefit** across operating points over threshold probability ranges to derive recommended operating points and **Δt** per case. **Weekly SPC** monitors drift (elongating **TeffT_\mathrm{eff}Teff**, declining **PresP_\mathrm{res}Pres**, skewed **FPR/FNR**). **Calibration (Brier/ECE)** assesses prediction–observation agreement and informs falsification patches (Δt, notification counts, UI wording).

Application follows **proposal → human approval**, preserving **shared decision-making**.

All logs and settings are sealed as **evidence ZIP + SHA256**, linked to the **pre-registration ID** and **SSOT (GitHub SHA)** for public reproducibility.

Primary outcomes reported are **≤10-second attainment rate**, **cross-Tier penetration failure rate**, and **Net Benefit dominance**, yielding practical operating-point ranges (Quiet/Standard/Aggressive) that support **expected harm minimization**.

10. Related Work

SiriusA Core (Pillars)

Spanning time-constrained HRI/HF, human-in-the-loop AI ethics, and gerontechnology usability, SiriusA operationalizes **protection of life** through an enforceable rule set centered on the **10-second ritual** and **shared decision-making**. Its key differentiator is that humans **select** both the **operating point (Quiet/Standard/Aggressive)** and the **approval window (Δt)**, thereby minimizing **fatal delay** in real-world use.

Objective

Synthesize prior work and make explicit SiriusA's novelty: jointly optimizing **time** (**TeffT_\mathrm{eff}Teff / PresP_\mathrm{res}Pres**) while upholding **ethical boundaries**.

1. **HRI/HF:** Prior art covers warning design, double checks, and multimodal (haptic/visual/auditory) cues, but rarely provides a framework that links and co-optimizes TeffT_\mathrm{eff}Teff and PresP_\mathrm{res}Pres.
2. **AI ethics:** "Human-in-the-loop" is often conceptual. SiriusA codifies it as policy by mandating **two-step confirmation** and a **revoke code**, and by excluding automatic actions.
3. **Gerontechnology:** Extensive work on cognitive load exists, yet the integration of **family multisig** with the **10-second ritual**—a "family consensus × time optimization" mechanism—remains under-specified.

a) **Linked metrics:** Teff=Δtset+Pres−1T_\mathrm{eff} = \Delta t_\mathrm{set} + P_\mathrm{res}^{-1}Teff=Δtset+Pres−1; **FPR/FNR**; **Net Benefit**

b) **Methods: decision curve analysis**; **calibration (Brier/ECE)**; **weekly SPC**

c) **Records: non-PII KPIs** with **evidence ZIP + SHA256** for recomputability

Pitfalls (≤2)

- Treating related work's "ideal UI" assumptions as given causes real-world mismatch → require operational rule sets.
- Ethics-only treatments that omit time optimization leave **fatal delay** unresolved.

HRI/HF under time pressure supports alerts, staged confirmations, and modality integration, yet seldom centers on **operational co-optimization** of TeffT_\mathrm{eff}Teff and PresP_\mathrm{res}Pres. SiriusA locks in a human-choice structure for **Δt** and the **operating point (Quiet/Standard/Aggressive)**, addressing **fatal delay** under **expected harm minimization**. Whereas "human-in-the-loop" ethics is typically abstract, SiriusA specifies it as an operational regime: enforce **two-step confirmation** and a **revoke code**, and—even for **direct-call phrases (119/110)** as a **WAIT48h** exception—retain **two-step**. In gerontechnology, despite rich studies on cognitive burden and usability, few systems formalize "family reachability" via **family multisig** while guaranteeing decision reversibility through the **10-second ritual**. By preserving **non-PII KPIs** with **evidence ZIP + SHA256** and evaluating TeffT_\mathrm{eff}Teff, PresP_\mathrm{res}Pres, **FPR/FNR**, and **Net Benefit** with **weekly SPC**, **calibration (Brier/ECE)**, and **decision curve analysis**, this work bridges prior knowledge into a combined **rule-set + measurement** paradigm.

To our knowledge, no prior HRI/assistive-AI work integrates ZK-based qualification proofs with duress-aware Δt control and family multisig for human-centered safety.

---

11. Limitations, Risks, and Legal Considerations

Guide: Do not assume zero risk. Always report residual **FPR/FNR** together with **(T_\mathrm{eff} / P_\mathrm{res})**, and keep selection under human control.

SiriusA Core (Pillars)

There is no such thing as 100% safety. By making residual **FPR/FNR** and expected harm visible, SiriusA maximizes **protection of life** while preserving **shared decision-making**.

Objective

State the limits of the **non-medical** scope and translate **expected harm minimization** into operational rules that maintain responsibility boundaries and deliver accountability through evidence.

1. **Residual-risk management:** Display **FPR/FNR** alongside **effective time ((T_\mathrm{eff}))** and **responsiveness ((P_\mathrm{res}))**; humans choose the **operating point (Quiet/Standard/Aggressive)** and the **approval window (Δt)**.
2. **Legal boundary:** Medical judgments are delegated to external parties (119/physicians). The system forbids **automatic transmission, payment, and reporting**.
3. **Accountability: Non-PII KPIs** plus **evidence ZIP + SHA256** ensure that decisions and revocations are reproducible.

a) **Metrics:** FPR, FNR, (T_\mathrm{eff}), (P_\mathrm{res}), False Alarm Burden (person-hours)

b) **Methods:** weekly SPC; **calibration (Brier/ECE)**; **decision curve analysis**

c) **Records:** two-step timestamps; **revocation latency**; family multisig (**k/n, Δt**) logs

Pitfalls (≤2)

- Assuming "zero risk" by default breaks operations; thresholds and the **operating point (Quiet/Standard/Aggressive)** must be public and explicit.
- Always-on monitoring violates ethical/legal limits; **Co-op** must remain **Verifier-only** and **trigger-only**.

We clarify the limits of a **non-medical** system. SiriusA lowers **fatal delay**, but it cannot eliminate **FPR** and **FNR**. Residual risk is therefore presented together with **(T_\mathrm{eff} / P_\mathrm{res})**, and humans **select the operating point (Quiet/Standard/Aggressive)** and **approval window (Δt)**. Medical decisions are routed to **direct-call phrases (119/110)** or physicians; even under **WAIT48h** exceptions, **two-step confirmation** and the **revoke code** are retained. The system never performs **automatic transmission, payment, or reporting**.

We fix responsibility as **AI proposes; humans decide**. We record **two-step** and **revoke** timestamps, family multisig (**k/n, Δt**), approver Tier, and the chosen

operating point as **non-PII KPIs**, sealed in an **evidence ZIP + SHA256**. **Weekly SPC** detects drift; **calibration (Brier/ECE)** checks prediction–outcome agreement; and **decision curve analysis** compares net benefit across operating points. Changes are applied via **proposal → human approval**, ensuring accountability through recomputation.

In deployment, reconcile interests among families, enterprises, co-ops, and municipalities; to avoid surveillance normalization, keep **Co-op** strictly **Verifier-only / trigger-only**. Thus, expected harm is minimized while human dignity and legal–ethical boundaries are preserved.

12. Conclusion and Future Work

SiriusA Core (Pillars)

Our purpose is **protection of life**. SiriusA centers the **10-second ritual** and **shared decision-making**, reducing **fatal delay** by optimizing **TeffT_\mathrm{eff}Teff** and **PresP_\mathrm{res}Pres**. The next step is to bridge this operational rule set to real-world deployment through **PoCs** and **standardization**.

Objective

Pursue a staged plan—**D0 → D7 → D30 → D90**—to test hypotheses and advance standardization under **secret thresholds** and **public evidence**.

1. **Roadmap: D0** (this paper) → **D7** (prepare small PoC) → **D30** (run PoC) → **D90** (scale-up & audit).
2. **PoC:** Verify the hypothesis that each household can optimize its **operating point (Quiet/Standard/Aggressive)** and **approval window (Δt)**.
3. **Standardization:** Keep thresholds confidential; publish **non-PII KPIs** and **evidence ZIP + SHA256**; uniquely bind releases via **SSOT (GitHub SHA)**.

a) **KPIs:** TeffT_\mathrm{eff}Teff, PresP_\mathrm{res}Pres, FPR, FNR, Net Benefit, calibration (Brier/ECE)

b) **Methods:** decision curve analysis; weekly SPC (drift detection)

c) **Operations:** Pre-register and enable recomputation for **two-step, revoke, family multisig (k/n, Δt)** logs

Pitfalls (≤2)

- Overreliance on auto-adaptation erodes human approval.
- Publishing too many indicators risks privacy leakage (preserve the **minimal set**).

D0 has specified a framework to minimize **fatal delay** by ensuring **humans choose** both the **operating point** and the **approval window (Δt)**, anchored by:

Teff=Δtset+Pres−1T_\mathrm{eff} = \Delta t_\mathrm{set} + P_\mathrm{res}^{-1}Teff=Δtset+Pres−1

At **D7**, we will finalize the PoC plan and prepare small studies across the four cases (fraud / disaster / fall / family communication) × **operating point (Quiet/Standard/Aggressive)**.

At **D30**, we will execute the PoC, collect **non-PII KPIs** (Decision Time, PresP_\mathrm{res}Pres, FPR, FNR, Net Benefit, Brier/ECE), and derive recommended **operating points** and **Δt** using **weekly SPC** and **decision curve analysis**.

By **D90**, we will establish expansion and an audit framework, fixing the principle of **threshold secrecy** with **evidence disclosure** via **evidence ZIP + SHA256** and **SSOT (GitHub SHA)**.

Open research directions include: optimizing any **automatic adaptation** of the **operating point** behind a **human-approval gate** (**two-step + revoke code**); generalizing **family multisig (k/n)** and time-of-day presets; and field-validating **expected harm minimization**, including **False Alarm Burden (person-hours)**.

SiriusA moves standardization—**terminology, measurement, evidence**—to the forefront and connects, step by step, to social deployment.

Integrating a Zero-Knowledge family confirmation layer shows that fatal-decision prevention is achievable without revealing identities: authorization is proved, not exposed, while two-step/revoke and Δt-based controls remain intact.

13.Epilogue — The Foundational Propositions of SiriusA

1. Architectural Ethics

This work treats responsibility not as a fixed attribution but as a **temporal vector** that moves through the decision process.

SiriusA implements **operational ethics** through **two-step confirmation**, a **revoke code**, and maintenance of the **WAIT48h exception**, suppressing **fatal delay** via the **10-second ritual**.

Humans retain sovereignty by selecting the **operating point (Quiet/Standard/Aggressive)** and managing the **approval window (Δt)**.

Consequently, accountability extends beyond "who decided" to include **when and how** the decision occurred—auditable through **evidence ZIP + SHA256**.

2. Multi-Model Convergence Principle

The reliability of complex hypotheses is determined not by the likelihood of a single model but by the **agreement among heterogeneous AI systems**.

In practice, an **agreement threshold (τ)**—for example, **Cohen's κ** or **mean Jensen–Shannon divergence (JSD)**—is **pre-registered**.

If agreement falls below τ, the **operating point** is **shifted one level toward Quiet** to reinforce **two-step confirmation**.

Model names, versions, and prompts are archived in the appendix, while **final approval always rests with a human**.

This structure combines **verifiability** with **reproducibility**.

3. Self-Amplifying Structure (Time as an Ally)

As AI capabilities and diversity grow, the **search space for convergence verification** expands, causing the theoretical framework to **strengthen monotonically over time**.

Analogous to trust accumulation in distributed systems, this approach avoids dependence on any **single point of failure**.

Using **weekly SPC (statistical process control)** and **calibration (Brier/ECE)**, SiriusA continuously optimizes the balance between **responsiveness (P_res)** and **safety margin** in operation.

4. Synthesis and Implications

SiriusA unifies **time-constrained operational ethics**, **multi-model convergence**, and **temporal self-strengthening**, proposing a paradigm whose **objective function is safe action within 10 seconds**.

Grounded in **family multisig (k/n, Δt)** and **non-PII KPIs**, it offers a reproducible framework for social deployment **without altering its methods or outcomes**.

Future cross-domain PoCs will empirically calibrate the **agreement threshold (τ)** and **human responsiveness distributions**, advancing the principle of **standardization through confidential thresholds and public evidence**.

Reproducibility & SSOT
- Repository:decision-os-paper
- PDF:D0_EN_Master_.pdf
- **SSOT commit:**840c85de344f5dd197dc5122ad9f4ba452f2970d

This PDF corresponds exactly to this commit. Revisions will be tagged; figures/tables are reproducible from this SSOT.