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

Shinichi Nagata

Abstract

This paper presents Decision-OS V5, code-named SiriusA, a policy-level confirmation layer for human–AI shared decision-making in safety-critical, non-medical domains. SiriusA treats the protection of human life as its primary objective and enforces human final consent through two-step confirmation and an explicit revoke path. We decompose effective time as $T_{\mathrm{eff}} = \Delta t_{\mathrm{set}} + P_{\mathrm{res}}^{-1}$, separating the approval window Δt from human responsiveness P_{res} as independently controllable factors.

The framework integrates family multisig (m-of-k, Δt) and a Zero-Knowledge confirmation layer (310/320) that verifies authorization without exposing personally identifiable information. Safety is operationalized via auditable state transitions, non-PII KPIs, and evidence packaging (ZIP + SHA256), while explicitly forbidding automatic transmission, payment, or reporting. Although this work focuses on life-protective decisions, the confirmation-layer architecture is designed to be extensible to broader classes of irreversible actions without altering its responsibility boundaries. Here, "non-medical" excludes clinical diagnosis or treatment decisions, while including safety-related scenarios such as fraud prevention, disaster response, and family coordination.Final consent and responsibility always remain with humans.

Index Terms

human-in-the-loop, confirmation layer, zero-knowledge proof, decision safety, duress-aware control, two-step confirmation, auditability

I. Introduction

A. Motivation

Digital decision-making increasingly involves irreversible actions (e.g., cryptographic signing, financial transfers, and account recovery). In high-pressure situations, users may be forced, confused, or socially engineered into authorizing harmful actions. Existing authentication factors reduce unauthorized access but do not sufficiently address *coerced* or *panic-driven* authorization.

B. Problem Statement

We target the following failure mode: a legitimate user, operating under duress or acute confusion, performs a valid signature that triggers an irreversible outcome. The core challenge is to introduce a confirmation layer that (i) reduces harm under duress, (ii) preserves usability for legitimate actions, and (iii) provides verifiable audit evidence without exposing sensitive personal information.

C. Contributions

This paper proposes **Decision-OS V5 (SiriusA)**, a zero-knowledge confirmation layer for life-protective decision execution. Our main contributions are:

- A **confirmation layer** separating intent from execution via approve/revoke/execute/hold states.
- A **duress-aware control** mechanism using a risk signal (e.g., `duress_score`) and a time window Δt to delay or block execution when necessary.
- A **two-step confirmation** protocol with **revoke code** and optional **family multisig (m-of-k)** for high-risk actions.
- An **evidence and audit** design using non-PII logs and evidence packaging (ZIP + SHA256) to support verification.
- A **panic mode** operational protocol: stop, preserve evidence, notify trusted parties, and use direct emergency phrases for 119/110 (exceptions to waiting).

D. Paper Organization

The remainder of this paper presents the system overview, protocol and state transitions, architecture, time optimization, interaction design, family multisig, operational audit, evaluation, and concluding discussion.

II. System Overview

A. Design Goal

Decision-OS V5 introduces a confirmation layer that separates *intent* from *execution*. The primary goal is to reduce irreversible harm caused by coerced, confused, or panic-driven authorization, while preserving usability for legitimate actions.

B. Actors

The system involves the following actors:

- **User**: the legitimate account holder.
- **Execution Target**: wallet, account, or system receiving the final authorization.
- **Confirmation Layer**: an intermediate control layer that can approve, delay, revoke, or block execution.
- **Trusted Parties (optional)**: family members or designated family verifiers participating via multisig.

C. Confirmation States

All high-risk actions are mediated by the confirmation layer, which maintains the following states:

- **Approve**: intent is accepted but not yet executed.
- **Execute**: irreversible action is performed.
- **Hold**: execution is delayed within a time window Δt.
- **Revoke**: intent is cancelled using a revoke code or guardian intervention.

D. Execution Flow

A requested action does not directly trigger execution. Instead, it is routed through the confirmation layer, where additional checks, delays, or revocation paths may apply. This separation enables intervention under duress without exposing sensitive personal information.

III. Protocol and State Transitions

A. Risk Signals

The system monitors a set of risk signals that may indicate coercion, confusion, or panic-driven decision-making. These signals do not attempt to infer intent, but instead provide conservative triggers for intervention.

Examples include:

- abrupt urgency or threat cues,
- repeated confirmation failures,
- abnormal interaction timing,
- explicit distress or confusion indicators.

B. Duress Score

A composite risk indicator (`duress_score`) aggregates observed signals into a bounded value. The score is used solely for control flow decisions (e.g., delay or halt execution), and is not exposed as personal or behavioral profiling data.

C. State Machine

All protected actions follow a finite set of states:

- **Request**: user intent is declared.
- **Approve**: intent is accepted but not yet executed.
- **Hold**: execution is delayed within a time window Δt.
- **Execute**: irreversible action is performed.
- **Revoke**: intent is cancelled prior to execution.

D. Transition Rules

State transitions are governed by simple, auditable rules. Elevated risk signals may force transitions into **Hold** or **Revoke**, while low-risk conditions allow progression to **Execute**. This design ensures that intervention is possible without requiring probabilistic inference of user intent.

IV. Architecture Overview (SiriusA)

A. Roadmap

This chapter presents the SiriusA overview. Details appear in Section V (time optimization: T_{eff}, P_{res}), Section VI (HRI/UX: the 10-second ritual and the five-line UI), and Section VII (family multisig, $m - of - k, \Delta t$). Design principles are discussed later in the paper, followed by operational audit and evaluation sections.

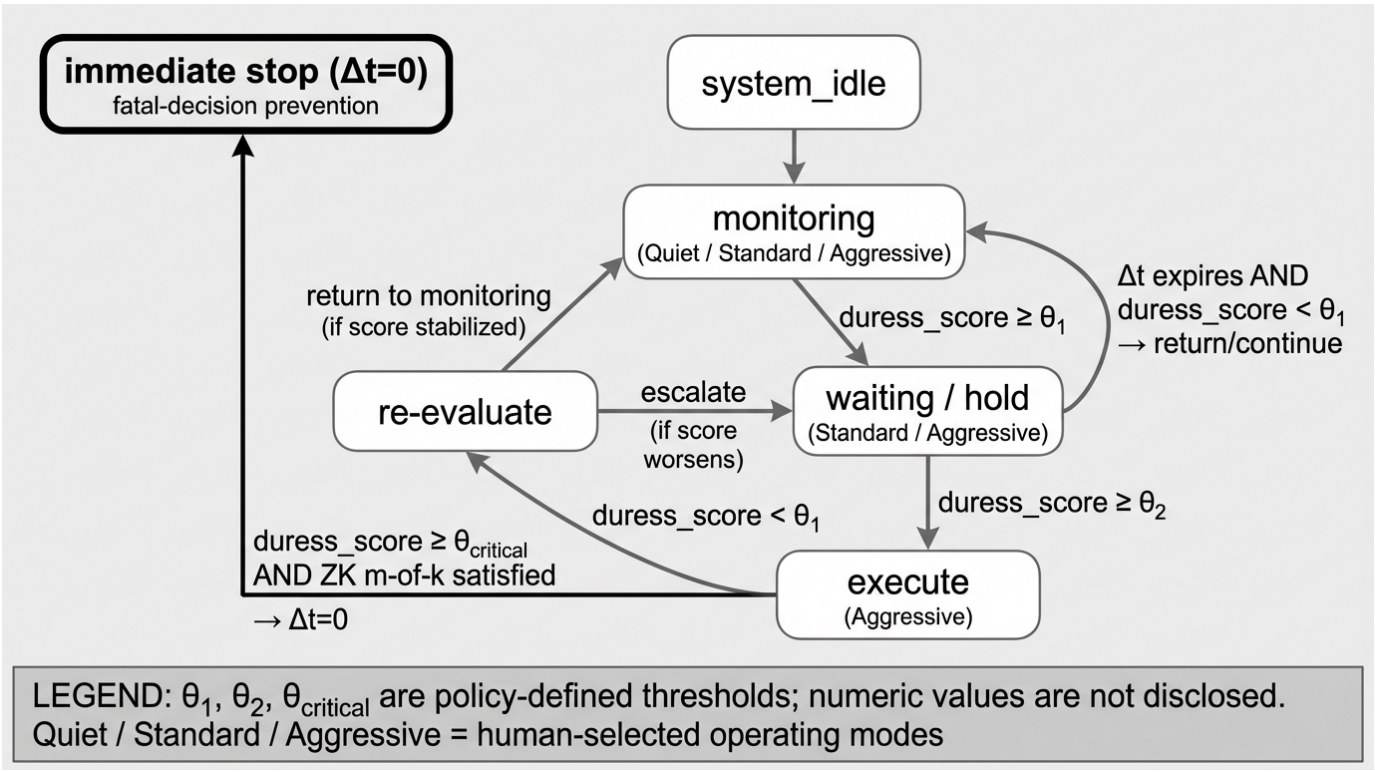

Fig. 1. State transitions controlled by duress_score thresholds, ZK m-of-k, and the approval window Δt.

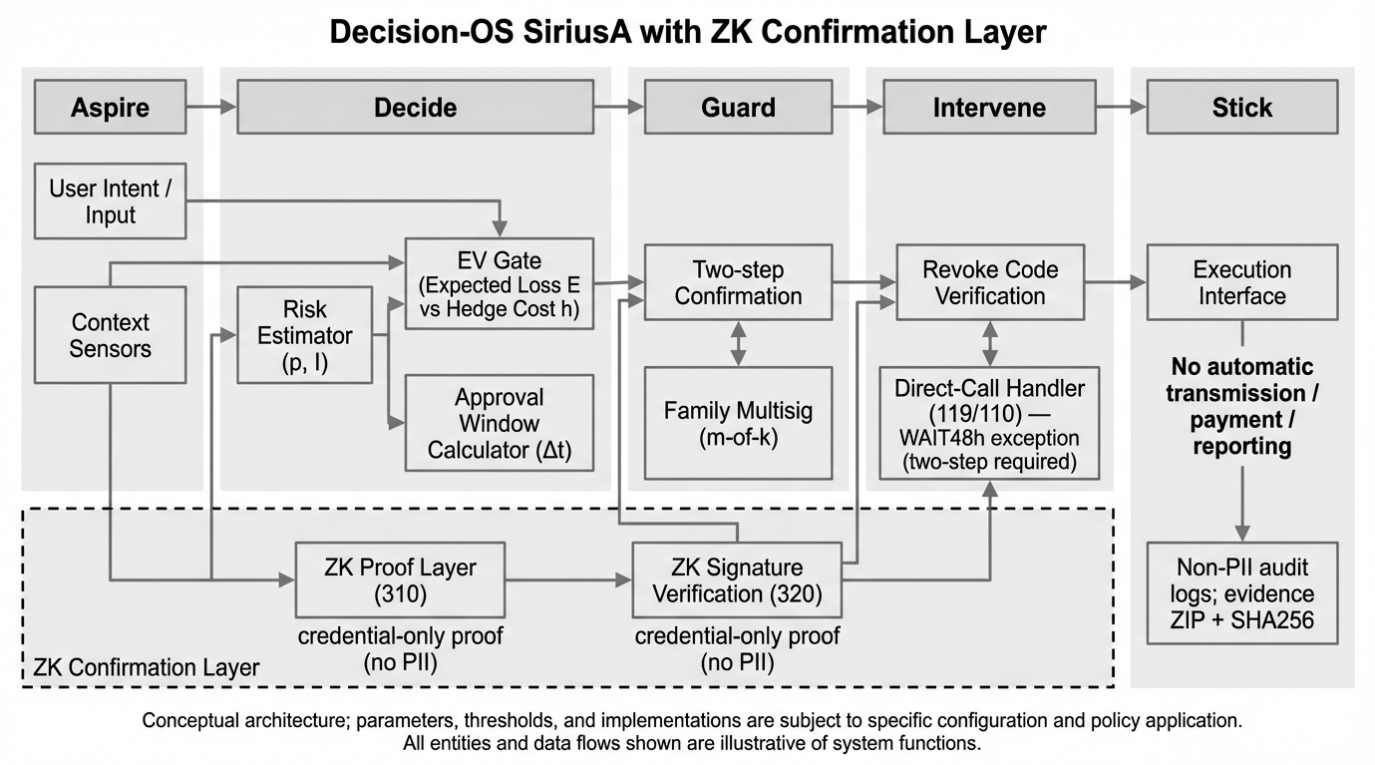

Fig. 2. Policy-level architecture of Decision-OS SiriusA with a ZK confirmation layer (310/320). (Note: numbers 310 and 320 denote conceptual module identifiers used consistently across figures.)

B. 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 T_{eff} and suppress false alarms. Evidence ZIP + SHA256 and signer-ID logs make operations auditable via non-PII KPIs.

C. 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 T_{eff}.
2) The operating point (Quiet/Standard/Aggressive) is selected by humans as a point on the receiver operating characteristic (ROC) curve, adapted by tuning P_{res} and Δt.
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) Ethical Boundaries (E-1–E-8).: The system enforces the following ethical boundaries:

- **E-1** No automatic execution without explicit human confirmation.
- **E-2** No automatic transmission, payment, or reporting.
- **E-3** Human final responsibility is always preserved.
- **E-4** Reversibility is required prior to execution (revoke path).
- **E-5** Non-PII operation and auditing are mandatory.
- **E-6** Mode switching requires explicit human action.
- **E-7** Emergency exceptions do not bypass two-step confirmation.
- **E-8** All actions are auditable via evidence ZIP + SHA256.

D. Figure, KPIs, and Logs

- **B**PMN/UML flow diagram (SiriusA loop around the perimeter). The core relation is:

$$T_{\text{eff}} = \Delta t_{\text{set}} + P_{\text{res}}^{-1}.$$

- **KPIs:** Decision Time, Adherence, FPR, FNR, Net Benefit, Brier, ECE.
- **Logs:** two-step confirmation, revoke code, timestamps for direct-call phrases (119/110) (WAIT48h exception), and signer IDs.

E. Pitfalls (2)

- Permitting automatic transmission, payment, or reporting risks bypassing human approval and collapsing responsibility under false alarms.
- Exceeding five UI lines degrades P_{res} and worsens T_{eff}.

F. Pipeline Details

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

$$T_{\text{eff}} = \Delta t_{\text{set}} + P_{\text{res}}^{-1}.$$

The approval window (Δt_{set}) is governed by two-step confirmation and family multisig ($m - of - k, \Delta t$); responsiveness (P_{res}) is raised by the five-line UI, direct-call phrases (119/110), the family one-liner, and routinized procedures.

TABLE I
DOMAIN-SPECIFIC Δt PRESETS (TABLE 1).

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	Adaptive	(time-of-day dependent)	Adaptive based on time-of-day / availability.

TABLE II
*

Note: Δt presets are operational ranges; policy thresholds (θ_1, θ_2, $\theta_{\mathrm{critical}}$) remain undisclosed.

A Quiet default prevents externalization of false alarms; Standard/Aggressive are selected explicitly when time is more valuable. The BPMN/UML flow diagram 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 the family multisig section and the BPMN/UML flow diagram.

V. TIME-OPTIMIZATION MODEL AND OPERATING-POINT SELECTION

At the center is effective time (T_{eff}), defined as:

$$T_{\mathrm{eff}} = \Delta t_{\mathrm{set}} + P_{\mathrm{res}}^{-1}. \tag{1}$$

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.

A. Objective

To present a mathematical time structure and an optimal operating-point model that minimizes T_{eff} while preserving human judgment.

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

Implementation flow and gate layout: see Section VII and Figure 1 (family multisig).

B. KPIs and Logs

KPIs: Decision Time, FPR, FNR, Net Benefit, Brier, ECE.
Logs: approval timestamps, Δt exceedance rates, two-step/revoke code/family ($m - of - k, \Delta t$) execution records.

C. Pitfalls ($\leq$2)

- Setting Δt too short $\rightarrow$ approval incomplete (FNR$\uparrow$); too long $\rightarrow$ worsens T_{eff}.
- Uncontrolled mode switching can raise false positives and undermine Quiet-mode stability.

D. Operational Notes

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 ($m - of - k, \Delta t$), 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.

VI. Interaction Design (HRI/HF)

A. Design Principles

The interaction layer is designed to reduce cognitive load under stress while preserving human agency. The core principles are:

- **Simplicity:** present only the minimum information required to decide.
- **Consistency:** identical layout across domains to build muscle memory.
- **Reversibility:** always provide a clear revoke path before execution.
- **Auditability:** every interaction leaves a non-PII trace.

B. The 10-Second Ritual

SiriusA introduces a fixed interaction window—the *10-second ritual*—to stabilize decision timing under pressure. Within this window, users review the action, confirm intent, or revoke. This ritual standardizes response time, improving predictability and safety.

C. Five-Line UI

All high-risk actions are rendered in five lines or fewer:

1) Action summary (what will happen).
2) Target (who/where the action affects).
3) Consequence (irreversibility and risk).
4) Confirmation method (two-step / revoke).
5) Next step or cancel instruction.

Limiting presentation to five lines reduces overload and increases responsiveness (P_{res}).

D. Family One-Liner

An optional *family one-liner* provides a concise, pre-agreed message that can be sent to trusted parties. This message communicates context without revealing sensitive data and supports rapid human verification.

E. Accessibility and HF Considerations

The interface accommodates human factors (HF) constraints:

- large typography and high contrast,
- reduced color dependence,
- minimal gestures,
- explicit language for emergency phrases (e.g., 119/110).

These choices aim to support elderly users and high-stress scenarios.

F. Operational Notes

Interaction design directly affects responsiveness (P_{res}). By constraining complexity and enforcing ritualized confirmation, SiriusA increases P_{res} without automation, keeping final consent human-controlled.

VII. Family Multisig and Gate Placement

A. Rationale

Certain actions carry elevated irreversible risk and benefit from shared human oversight. SiriusA supports an optional family multisig to distribute responsibility while preserving user agency.

B. Family Multisig Model

A family multisig is defined by parameters $(m - of - k, \Delta t)$:

- n: number of designated trusted parties.
- k: minimum approvals required for execution.
- Δt: approval window aligned with the operating point.

Execution proceeds only after k approvals are collected within Δt; otherwise the action remains on hold or is revoked.

C. Integration with Two-Step Confirmation

Family multisig complements two-step confirmation:

- **Step 1:** user approval creates an intent record.
- **Step 2:** family approvals (if enabled) finalize execution.

A revoke code can cancel the most recent approval at any time before execution.

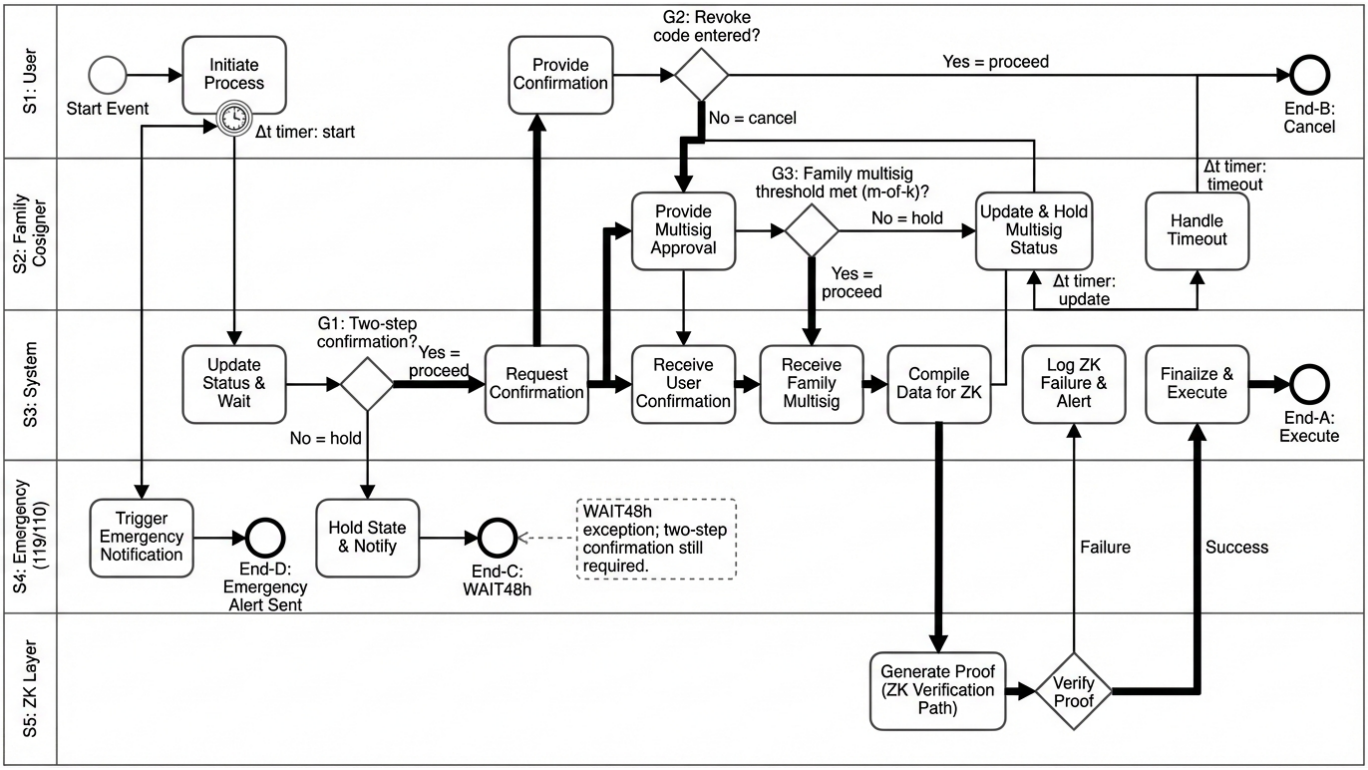

Fig. 3. BPMN-like confirmation flow with two-step confirmation, revoke, family multisig (m-of-k), and ZK verification.

D. Gate Placement

Gate placement determines where checks are enforced in the execution pipeline:

- **Pre-gate:** before irreversible actions (default).
- **Mid-gate:** between intent and execution for high-risk domains.
- **Post-gate:** audit-only logging after execution.

SiriusA defaults to pre-gate placement to maximize intervention opportunity.

E. the BPMN/UML flow diagram

The BPMN/UML flow diagram depicts the SiriusA loop around the perimeter: trigger detection, mode management, ritual UI, family one-liner, two-step confirmation, execution, and signed logging. An outer weekly SPC loop detects drift and proposes threshold updates, subject to explicit human approval.

F. Operational Notes

Family multisig is optional and context-dependent. It is recommended for financial transfers, account recovery, and external reporting, and may be disabled for time-critical emergencies where delay would increase harm.

VIII. Safety Verification and Operational Audit

A. Verification Scope

Safety verification focuses on whether the system preserves human final consent, prevents irreversible harm under duress, and produces verifiable evidence without collecting personal data.

B. Evidence Packaging

All protected actions generate an *evidence package*:

- **Logs:** timestamps, state transitions, approvals, revocations.
- **Artifacts:** hashes, signer identifiers, mode selections.
- **Bundle:** evidence ZIP with SHA256 digest.

The package is immutable once sealed and can be verified independently.

C. Non-PII KPIs

Verification relies on non-PII KPIs to avoid behavioral profiling:

- decision latency (T_{eff}),
- false positive / false negative rates (FPR/FNR),
- adherence to operating points,
- revoke utilization rate.

D. Audit Procedure

Audits may be performed internally or by third parties:

1) verify the integrity of the evidence ZIP via SHA256,
2) check state transitions against protocol rules,
3) review KPI thresholds and deviations,
4) confirm that no execution bypassed human approval.

E. Operational Boundaries

Audit rules enforce ethical and responsibility boundaries:

- no automatic execution without human confirmation,
- no silent mode switching,
- no export of raw interaction data.

Violations are detectable through missing or inconsistent evidence.

F. Continuous Monitoring

A weekly statistical process control (SPC) loop monitors drift in KPIs. Threshold updates are proposed automatically but applied only after explicit human approval, preserving accountability.

IX. Evaluation Plan and Metrics

A. Evaluation Objectives

The evaluation assesses whether Decision-OS V5 (SiriusA) reduces irreversible harm while maintaining usability and human agency. The focus is on operational effectiveness rather than predictive accuracy.

B. Primary Metrics

The following key performance indicators (KPIs) are used:

- **Decision Time (T_{eff}):** effective time to safe execution.
- **False Positive Rate (FPR):** unnecessary interventions.
- **False Negative Rate (FNR):** missed harmful actions.
- **Net Benefit:** expected harm reduction minus intervention cost.

C. Secondary Metrics

Additional metrics provide calibration and robustness signals:

- Brier score and expected calibration error (ECE),
- adherence to selected operating points,
- revoke utilization and recovery success rate.

D. Evaluation Protocols

Evaluation is conducted through controlled simulations and pilot deployments:

1) define domain-specific scenarios (fraud, disaster, fall, family),
2) deploy the confirmation layer with fixed thresholds,
3) collect non-PII logs and evidence packages,
4) compute KPIs and compare against baseline workflows.

E. Ethical Considerations

Evaluation protocols avoid deception, profiling, or coercive testing. Participants retain full control and may withdraw at any time. All assessments respect local emergency procedures and legal requirements.

F. Limitations

Metrics capture operational performance but do not guarantee prevention of all harm. Results may vary across domains and user populations, and long-term effects require continued observation.

G. Related Work

Related standards for strong authentication and lifecycle management include the NIST 800-63 suite and the W3C WebAuthn recommendation. Prior work on duress signaling motivates revoke- and hold-oriented designs, while multisignature schemes provide a basis for optional m-of-k family co-approval.[1], [2], [3], [4]

H. Limitations

Decision-OS V5 (SiriusA) prioritizes human final consent and auditability. Accordingly, it does not guarantee prevention of all harm. Key limitations include:

- **Context variability:** Operating points and Δt presets may require domain- and culture-specific tuning.
- **Residual false negatives:** Some coercive scenarios may evade detection, especially when signals are intentionally minimized.
- **Deployment constraints:** Effectiveness depends on correct integration with wallets, accounts, and notification channels.

I. Discussion

Prior approaches focus on authentication factors, anomaly detection, or automation. In contrast, SiriusA emphasizes a *confirmation layer* that separates intent from execution, preserves reversibility via revoke codes, and produces non-PII evidence for audit. This positioning complements—rather than replaces—existing security controls.

J. Conclusion

We presented **Decision-OS V5 (SiriusA)**, a zero-knowledge confirmation layer designed to protect life by reducing irreversible harm under duress, confusion, or panic. By decomposing effective time

$$T_{\text{eff}} = \Delta t_{\text{set}} + P_{\text{res}}^{-1},$$

the system enables humans to select operating points that balance responsiveness and safety. SiriusA integrates a ritualized interaction (10-second ritual, five-line UI), two-step confirmation with revoke codes, optional family multisig, and auditable evidence packaging (ZIP + SHA256) without collecting personal data.

Across architecture, protocol, interaction design, and audit, the system maintains clear responsibility boundaries and verifiability. Future work includes longitudinal evaluation across domains, refinement of presets, and broader ecosystem integration. Ultimately, SiriusA demonstrates that safety and usability can be aligned through human-centered, auditable confirmation rather than opaque automation.

Disclosure and Appendix

Scope and Non-Claims

Decision-OS V5 (SiriusA) is a conceptual and architectural framework. It does not claim to be a complete implementation, a predictive model, or an autonomous decision-making system. Final judgment and responsibility always remain with humans.

Ethical Boundaries

The system enforces the following ethical boundaries:

- No automatic execution without explicit human confirmation.
- No hidden optimization or silent mode switching.
- No collection or inference of personal behavioral profiles.
- No delegation of moral or legal responsibility to the system.

Assumptions

The framework assumes:

- availability of trusted execution and logging environments,
- user-defined trusted parties for optional family multisig,
- compliance with local emergency procedures (e.g., 119/110).

Appendix: Four Propositions

The following propositions summarize the conceptual position of SiriusA. They are not empirical claims but design-oriented theses.

1) **Confirmation over Automation:** Safety in irreversible decisions is better achieved by inserting a human-centered confirmation layer than by increasing automation.
2) **Time Decomposition:** Effective time to harm can be reduced by decomposing delay into an approval window (Δt) and human responsiveness (P_{res}), rather than by minimizing latency alone.
3) **Reversibility as a First-Class Property:** The presence of an explicit revoke path changes human behavior and reduces panic-driven errors, even if revocation is rarely used.
4) **Auditability without Surveillance:** Verifiable safety can be achieved through non-PII evidence and cryptographic integrity, without continuous monitoring of users.

Reproducibility & Genesis Anchor

- Repository (SSOT): `decision-os-paper`
- SSOT commit (Git): `840c85de344f5dd197dc5122ad9f4ba452f2970d`
- DOI (V5 v1, Genesis): `10.5281/zenodo.17480645`
- This PDF corresponds exactly to the SSOT commit above. Revisions will be tagged.

REFERENCES

[1] P. A. Grassi, M. E. Garcia, and J. L. Fenton, "Digital Identity Guidelines," National Institute of Standards and Technology, Tech. Rep. NIST Special Publication 800-63-3, 2017. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-63-3

[2] W3C Web Authentication Working Group, "Web Authentication: An API for accessing Public Key Credentials, Level 2," W3C Recommendation, 2021. [Online]. Available: https://www.w3.org/TR/webauthn-2/

[3] J. Clark and U. Hengartner, "Panic Passwords: Authenticating under Duress," in *USENIX Workshop on Hot Topics in Security (HotSec)*, 2008. [Online]. Available: https://www.usenix.org/event/hotsec08/tech/full_papers/clark/clark.pdf

[4] M. Fukumitsu *et al.*, "A Tightly-Secure Lattice-Based Multisignature," in *ACM Conference on Computer and Communications Security (CCS)*, 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3327958.3329542