D0_JP_Master

[SiriusA 初期条件（Synced Block｜編集は同期元のみ）]

▼ 冒頭：責任の宣誓（Author Responsibility Declaration）

責任の宣誓（Author Responsibility Declaration）。 本稿における**研究設計、主張、評価指標、作動点選択（Quiet/Standard/Aggressive）と10秒儀式の定義、ならびに掲載の最終判断と責任**は**筆者が単独で負う**。AI（**OpenAI GPT-5 Thinking：primary／Google Gemini Ultra DeepResearch：secondary**）は**対案生成・言語整形**に用いたが、**全てのaccept/rejectと編集は人である筆者**が実施し、**数値・方法の改変は行っていない**。本研究は**非医療**の範囲で、**自動送信・決済・通報なし／two-step confirmation／revoke code**の原則を守る。**期待被害最小化**を北極星とし、**証跡ZIP＋SHA256**と**non-PII KPIs、weekly SPC／calibration（Brier/ECE）で再現性と説明責任を維持する。〔SSOT: decision-os-paper/<commit-SHA>〕

▼

▶ 1. 序論（目的・貢献・Core Question）

【主張の柱（SiriusA Core）】
生命を最上位目的に置き、致死的遅延（fatal delay）を最小化する運用OSを提示する。

10秒内の行動支援を人–AIの共同意思決定として定式化する。

貢献：①時間最適化式 \(T_{\mathrm{eff}}\) とドメイン別Δt、②家庭マルチシグ（k/n, Δt）統合、③E-1〜E-8の実装規約、④KPI群と監査設計。

【定義（本章のみ）】
SiriusA：aspirational design architecture（人の尊厳と自律を守る理想志向の設計層）。

North Star：objective function（人の決定権を損なわず、持続的便益を最大化する指標）。

【目的】
protection of life を最上位に、10-second ritual と two-step confirmation で人

主導を保持しつつ fatal delay を最小化する。

1. 有効時間 \(T_{\mathrm{eff}}=\Delta t_{\mathrm{set}}+P_{\mathrm{res}}^{-1}\) を明示し、遅延要因を分解・最適化する。
2. family multisig（k/n, Δt）で承認と時間窓を統合し、安全運用を標準化する。
3. non-PII KPIs と evidence ZIP + SHA256、weekly SPC により運用を監査・較正（Brier/ECE）する。

a) 式：\(T_{\mathrm{eff}}=\Delta t_{\mathrm{set}}+P_{\mathrm{res}}^{-1}\)、図1：家庭マルチシグのBPMN/UML。

b) KPI：Adherence／Decision Time／FPR／FNR／Net Benefit／Brier／ECE。

c) ログ：two-step・revoke code・署名者ID（直通語119/110はWAIT48h例外だが二重確認は維持）。

落とし穴（≤2）
・自動送信/決済/通報は採用しない（責任分界と誤発報抑制のため）。
・作動点（Quiet/Standard/Aggressive）は人が選ぶ。閾値の自動提案は人の承認必須。

本稿は protection of life を北極星とし、非医療の安全クリティカル領域（詐欺・災害・転倒・家族連絡）における人–AIの shared decision-making を10秒内の行動単位として運用化する。致死的遅延（fatal delay）を縮減するため、意思決定の有効時間を \(T_{\mathrm{eff}}=\Delta t_{\mathrm{set}}+P_{\mathrm{res}}^{-1}\) で定式化し、承認窓（approval window, Δt）と反応性（responsiveness, \(P_{\mathrm{res}}\)）の双方を設計変数とみなす。アーキテクチャは人の最終同意を中核に据え、two-step confirmation と revoke code を通過しない自動送信/決済/通報を排する。直通語（119/110）は WAIT48h の例外とするが、二重確認は維持する。家庭マルチシグ（family multisig, k/n, Δt）は単身1/1フォールバックを含む承認構造であり、ROC 上の作動点（operating point）の選択を家庭側の価値関数に委ねる。評価は non-PII KPIs（Adherence, Decision Time, FPR/FNR, Net Benefit,

Brier/ECE）を用い、evidence ZIP + SHA256 により改ざん耐性を持つ証跡を公開する。weekly SPC でドリフトを監視し、decision curve analysis と較正指標で実効価値を検証する。これらを通じて、SiriusA（aspirational design）と North Star（objective function）の二層を、現実の運用規約として結合し、期待被害最小化（expected harm minimization）に資する“使える”標準を提示する。

案内：全体像は §4 アーキテクチャ概観、定義と運用境界は §2 問題設定とスコープ を参照。

SSOT：decision-os-paper / commit SHA= TODO

▼ ▶ 2. 問題設定とスコープ（非医療・安全クリティカル）

【主張の柱（SiriusA Core）】
本稿のスコープは、医療行為を除く安全クリティカル領域（詐欺・災害・転倒・家族連絡）に限定する。

AIは人の判断を代替しない。two-step confirmation と revoke code によって、人の承認が保持される。

致死的遅延（fatal delay）を「危険検知から行動完了までの全時間」と定義し、これを最小化する設計境界を明示する。

【目的】

AIが自律的に判断しない非医療領域において、人–AI協働の境界と致死的遅延の構造を明確化する。

運用境界（固定）
本稿は非医療（non-medical）領域に限定し、既定の作動点は**Quiet**。**自動送信・自動決済・自動通報は行わない。**すべての外部行為は**two-step confirmation**を通過し、取消コード（revoke code）で可逆とする。直通語（119/110）は**WAIT48h**の例外だが、**二重確認**は維持する。（実装詳細は**E-1〜E-8（§3）**を参照。）

1. スコープを非医療に限定することで、法的・倫理的リスクを最小化しつつ、設計の再現性を確保する。

2. 「人が承認するまで出力を保留する」構造が、人の責任とAIの即応性を両立させる。
3. 致死的遅延を時間要素で分解することで、システム側と人側の改善ポイントを数理的に示せる。

a) 時間モデル：\(T_{\mathrm{eff}}=\Delta t_{\mathrm{set}}+P_{\mathrm{res}}^{-1}\)（式1を再利用）

b) KPI：Decision Time／FPR／FNR／Net Benefit（Expected harm minimization 指標）

c) ログ：承認時刻・取消コード入力・家庭マルチシグ(k/n, Δt)の承認率

落とし穴（≤2）
・スコープを曖昧にすると、AIの介入範囲が拡大し、人の責任分界が崩れる。
・「即応性」だけを重視すると、人承認を飛ばす設計に傾き、倫理境界（E-1〜E-8）を逸脱する。

本稿で扱う安全クリティカル領域は、生命に直接影響を与えるが医療行為には該当しない「詐欺・災害・転倒・家族連絡」の4カテゴリである。これらは人が最終行動者であることを前提に、AIが行動支援を行う領域である。システムの目的は代行ではなく補助であり、人の意思決定を高速化し、致死的遅延（fatal delay）を縮減することにある。致死的遅延は「危険の発生から行動完了までの経過時間」と定義され、その中にはAIの検知・提示遅延、人の認知・承認遅延、実行遅延の3要素が含まれる。このうちAIが最適化できるのは前者2項目のうちの提示部分までであり、最終的な承認と行動は人の責務とする。

SiriusAは、この分界線を「two-step confirmation」と「revoke code」によって可視化し、AIの自動送信・自動決済・自動通報を禁止する。家庭マルチシグ（family multisig, k/n, Δt）は、単身や家族構成に応じて承認閾値と時間窓（approval window）を調整し、緊急時の人承認を支援する枠組みである。これにより、AIの即応性（responsiveness, \(P_{\mathrm{res}}\)）を維持しながら、誤警報率（FPR）と見逃し率（FNR）の最適点を人が選ぶ設計が可能になる。これが本稿のスコープ設定の中核であり、以後の章ではこの時間分解モデルを前提として、E-1〜E-8の倫理設計原則と作動点選択を論じる。

案内：アーキテクチャ全体像は §4 を参照。

SSOT：decision-os-paper / commit SHA= TODO

▼ ▶ 3. デザイン原則と倫理境界（E-1〜E-8）

案内：本章は原則（E-1〜E-8）の定義のみ。指標と手法は **§9** を参照。

【主張の柱（SiriusA Core）】

人の最終承認を中心に据えた eight ethics (E-1〜E-8) を、AI運用の設計原理として定義する。

自動化を抑制し、two-step confirmation と revoke code を制度化することで、人の選択権を保持する。

非個情KPI（non-PII KPIs）と証跡ZIP＋SHA256を導入し、倫理境界の遵守を監査可能な形で実装する。

【目的】

SiriusA のアーキテクチャにおける倫理境界（E-1〜E-8）を明示し、AI運用の限界と保証条件を定義する。

1. two-step confirmation／revoke code により、AIの誤作動・誤承認を防ぎ、人の最終責任を保つ。
2. Quiet／Standard／Aggressive の作動点を人が選択できる設計が、倫理的自律性を支える。
3. 非個情データのみに基づくKPI運用と証跡ZIP＋SHA256の公開が、説明責任と透明性を保証する。

a) 図：E-1〜E-8フレームワーク（倫理境界と責任分界）

b) **指標・解析：§9 を参照（KPI定義・calibration・decision curve・SPC を集約）**

c) ログ：revoke code履歴／承認時刻／ZIP＋SHA256署名リスト

落とし穴（≤2）
・E-1〜E-8のうちいずれかを緩めると、Quietモードでも倫理的破綻点が生じる。
・透明性確保を怠ると、監査不能なAI判断が生まれ、社会的信頼を失う。

———————

SiriusAの設計は、生命保護（protection of life）を目的としながらも、AIの自動化を最小限に抑える eight ethics（E-1〜E-8）を中核原理として定義する。E-1〜E-4は「人中心設計」、E-5〜E-8は「監査と透明性」に対応する。E-1：two-step confirmation、E-2：revoke code、E-3：Quiet／Standard／Aggressive の人選択モード、E-4：119/110 WAIT48h例外規則、E-5：non-PII KPIs、E-6：evidence ZIP＋SHA256、E-7：weekly SPCによる倫理ドリフト監視、E-8：監査可能ログと開示責任。これらを体系化することで、AIの行動はすべて「人の同意を前提とした協働行為」として定義される。

アーキテクチャは「禁止ではなく構造による制約」を採る。たとえば revoke code の二段階実装は、AIの出力がユーザー承認を通過しない限り外部に送信されない構造的制動である。また、Quiet／Standard／Aggressive の3モードは、期待被害最小化（expected harm minimization）の観点から人が選ぶ operating point として設計されている。これにより、反応性（responsiveness, \(P_{\mathrm{res}}\)と安全性のトレードオフを倫理的に管理できる。

さらに、SiriusAは non-PII KPIs を用いて個人情報を含まない形で運用評価を行い、証跡ZIP＋SHA256を公開することで、第三者が監査可能な透明性を担保する。この枠組みは単なる倫理宣言ではなく、監査可能性と再現性を備えた「実装規約（E-1〜E-8）」として機能する。これにより、AIと人が協働する全過程において、致死的遅延を最小化しつつ倫理境界を超えない実用的バランスが保たれる。

SSOT：decision-os-paper / commit SHA=TODO

▼ ▶ 4. アーキテクチャ概観（SiriusA：10秒儀式の運用OS）

ロードマップ：本章は SiriusA の**全体像**を示す。**§5＝時間最適化（T_eff, P_res）**、**§6＝HRI/UX（10秒儀式・5行UI）**、**§7＝家庭マルチシグ（k/n, Δt）**が各構成要素の詳説である。**設計原則は§3**、**運用監査は§8**、**評価・指標は§9**に集約する。

【主張の柱（SiriusA Core）】
10-second ritual を中核に、トリガ検出→モード管理→5行UI→家族一文→二重確認→ログ署名の直列パイプラインを定義する。

作動点（Quiet/Standard/Aggressive）は人が選択し、T_eff最小化と誤警報抑制を同時に満たす。

evidence ZIP + SHA256 と署名者IDログにより、非個情KPIで運用を監査可能にする。

【目的】
人主導の意思決定を毀損せずに、fatal delay を最小化する処理パイプラインと責任分界を提示する。

1. 直列パイプライン（Detect→Mode→UI→Family one-liner→Confirm→Log）が T_effの分解と最適化を可能にする。
2. 作動点はROC上の operating point として人が選ぶ設計であり、P_res とΔt の調整で適応させる。
3. 監査可能性（non-PII KPIs／ZIP+SHA256／署名者ID）が“実装規約”として倫理境界（E-1〜E-8）を支える。

a) 図1：BPMN/UML（外周にSiriusAループ）。式：\(T_{\mathrm{eff}}=\Delta t_{\mathrm{set}}+P_{\mathrm{res}}^{-1}\)。

b) KPI：Decision Time／Adherence／FPR／FNR／Net Benefit／Brier／ECE。

c) ログ：two-step／revoke code／直通語の使用時刻（119/110はWAIT48h例外）／signer IDs。

落とし穴（≤2）
・自動送信/決済/通報を許すと、人承認が抜け、誤発報時の責任分界が崩れる。
・UIが5行を超えると P_res が低下し、T_effが悪化する。

SiriusAは、10-second ritual を中心に据えた運用OSとして、直列パイプラインを採る。①Trigger Detection：詐欺・災害・転倒・家族連絡の各ドメインで軽量検知を行い、過検知はQuiet既定で吸収する。②Mode Management：Quiet/Standard/Aggressive の operating point を人が選択し、誤警報率（FPR）

と見逃し率（FNR）のトレードオフを家族の価値関数に合わせる。③Ritual UI（最大5行）：実行すべき行動を5行以内で提示し、家族一文を付すことで認知負荷を制御する。④Two-step Confirmation：送金・連絡・通報は必ず二重確認を通過し、取消コード（revoke code）で直近の承認を反転可能にする。直通語（119/110）はWAIT48hの例外だが、二重確認は維持する。⑤Evidence & Logging：non-PII KPIs を収集し、evidence ZIP + SHA256 と signer IDs で監査可能性を担保する。

この構造は \(T_{\mathrm{eff}}=\Delta t_{\mathrm{set}}+P_{\mathrm{res}}^{-1}\) を実装可能な形に分解する。承認窓 \(\Delta t_{\mathrm{set}}\) は二重確認と家庭マルチシグ（family multisig, k/n, Δt）で制御し、反応性 \(P_{\mathrm{res}}\) は5行UI・直通語・家族一文・既知の手順化により高める。Quiet既定により誤発報の外部化を防ぎ、Standard/Aggressive は時間価値の高い状況で人が明示的に選ぶ。全体はBPMN/UMLの図1で示し、外周に週次SPCのフィードバックループを配置して、ドリフト検知としきい値の自動提案（人の承認必須）を行う。これにより、人の最終同意・責任分界・監査可能性を損なわず、10秒内の shared decision-making を現実に運用できる。実運用時のゲート配置と分岐は **§7 図1参照**。

SSOT：decision-os-paper / commit SHA= TODO

▼ ▶ 5. 時間最適化モデルと作動点選択

表1：ドメイン別 Δt プリセット

ドメイン	既定のoperating point	推奨Δt(秒)	備考
詐欺	Quiet	90–180	誤承認コスト高。通知少・二重確認多。
災害	Standard	15–45	即応重視。定型の連絡テンプレ短縮。
転倒	Quiet	120–240	夜間はΔt延長、誤警報削減を優先。
家族連絡	Quiet	任意	柔軟運用。時間帯プリセット連動。

案内：本章は **§4** 構成要素「時間最適化」の詳説。境界は **§2**、指標・手法は **§9** を参照。

【主張の柱（SiriusA Core）】

有効時間 \(T_{\mathrm{eff}}=\Delta t_{\mathrm{set}}+P_{\mathrm{res}}^{-1}\) を中心に、承認窓（approval

window, Δt）と反応性（responsiveness, P_res）の両軸で致死的遅延（fatal delay）を最小化する。

Quiet／Standard／Aggressive の3作動点を人が選び、ROC上の operating point として人の価値関数に基づいて設定する。

各ドメイン（詐欺・災害・転倒・家族連絡）に応じて Δt と作動点の既定を分け、expected harm minimization を達成する。

【目的】

時間構造を数理化し、人の判断を保持したまま T_effを最小化する最適作動点モデルを提示する。

1. \(T_{\mathrm{eff}}\) は「設計変数」Δtと「行動変数」P_resの和であり、設計・運用の双方で制御可能である。
2. 作動点はROC上の operating point として人が選択し、Quiet既定により誤警報率（FPR）を抑制する。
3. ドメイン別時間価値を反映したΔtプリセットにより、FPR/FNR のトレードオフを現実的に管理する。

実行フローとゲート配置は **§7 図1（家庭マルチシグ）参照**。

a) 式：\(T_{\mathrm{eff}}=\Delta t_{\mathrm{set}}+P_{\mathrm{res}}^{-1}\)（式1再掲）／図：ROC曲線上の作動点

b) KPI：Decision Time／FPR／FNR／Net Benefit／Brier／ECE

c) ログ：承認時刻・Δt超過率・two-step／revoke code／family(k/n, Δt)実行記録

落とし穴（≤2）
・Δtを短くし過ぎると承認未達（FNR↑）、長くし過ぎるとTeff悪化。
・UIや通知量を増やしてP_resを高め過ぎると、認知負荷が上がりQuiet基準が崩れる。

SiriusAでは、致死的遅延（fatal delay）の最小化を目的に、有効時間 \(T_{\mathrm{eff}}=\Delta t_{\mathrm{set}}+P_{\mathrm{res}}^{-1}\) を定義す

る。ここで承認窓 \(\Delta t_{\mathrm{set}}\) は設計上の固定時間、反応性 \(P_{\mathrm{res}}\) は人とUIの相互作用で決まる可変要素である。Δtは二重確認（two-step confirmation）と家庭マルチシグ（family multisig, k/n, Δt）で制御され、P_res は5行UI、家族一文、通知パスの冗長化、直通語（119/110, WAIT48h例外）によって高められる。

Quiet／Standard／Aggressive の作動点は、人がROC上の operating point を選ぶ行為として定義される。Quietは過検知時の外部化を防ぎ、Standardは時間価値の高い状況で使用し、Aggressiveは致死リスクが目前にある際に限定的に選ばれる。作動点はAIが自動的に決めない。常に人が切り替え、revoke code により可逆性を担保する。

各ドメインの時間特性によりΔtの最適値は異なる。詐欺では誤承認コストが高く、Δt中程度・Quiet既定。災害は最短Δt・Standardで迅速化。転倒はΔt長め・Quietで安定。家族連絡は任意で柔軟設定とする。これにより、ドメイン別に expected harm minimization が可能となる。

週次SPCでは、non-PII KPIs（Decision Time, FPR, FNR, Brier/ECE）を用いてT_effのドリフトを監視し、ΔtとP_resの調整案を生成する。提案値は人が承認し、ZIP＋SHA256で証跡化する。AIは提案までを行い、実装は人が決定する。この「人が選ぶ最適点」構造により、時間最適化は倫理境界（E-1〜E-8）を維持したまま実用化される。これがSiriusAにおける時間最適化モデルの中核である。

SSOT：decision-os-paper / commit SHA= TODO

▼ ▶ 6. インタラクション設計（HRI / Human Factors）

案内：本章は **§4** 構成要素「HRI/UX」の詳説。境界は **§2**、指標・手法は **§9** を参照。

【主張の柱（SiriusA Core）】
「10秒儀式」を中核に、人が最終決定者の shared decision-making を固定。AIは提示・可視化・記録に限定し、protection of life に直結する fatal delay を最小化する。

【目的】
P_res を上げつつ T_eff を下げ、FPR/FNR を人が制御できる UI/UX を確立する。

1. 5行UI＝〔入力→再確認→承認→取消→静止〕で「10秒儀式」を実装し、revoke code を常時受付。
2. two-step confirmation＝1st入力→2秒クールダウン→2nd確定。assist は approval window (Δt) 残秒と operating point（Quiet/Standard/Aggressive）を明示。**§7 図1参照**。
3. family multisig（k/n, Δt）＋時間帯プリセットで Δt・通知本数・入力様式（音/光/触覚）を最適化。

a) KPI：Decision Time／P_res／FPR／FNR／Net Benefit／calibration（Brier/ECE）

b) 分析：decision curve analysis／weekly SPC（ドリフト検知）

c) ログ：non-PII KPIs＋evidence ZIP + SHA256（two-step時刻／revoke使用／family k/n達成率）

落とし穴（≤2）
・ガイダンス盛り過ぎ＝認知負荷↑→P_res↓（5行超は不可）。
・静的プリセット固定＝状況不適合（週次で閾値提案→人の承認で更新）。

本章は、人が最終承認者である前提を崩さず、T_eff と P_res の同時最適化を UI/UX で実装する。中核は「10秒儀式」。5行UIは〔入力→再確認→承認→取消→静止〕で構成し、全段に revoke code を常設する。two-step confirmation は 1st入力後に2秒の認知クールダウンを置き、2nd確定で完了。assist は approval window (Δt) の残秒と operating point（Quiet/Standard/Aggressive）を常時表示し、Quiet を既定とする（過検知の外部化を抑える）。Standard/Aggressive は状況依存で人が明示選択。

family multisig は k/n と Δt を持つ行動トークンとして扱い、家庭内2-of-2（スマホ＋ウォッチ/据置）を既定に、時間帯スロット（深夜/早朝/昼/夕/就寝前）で Δt・通知間隔・合図様式をプリセット化する。深夜は Δt を延長、昼は通知間隔を短縮するなど、実測に基づいて P_res を押し上げる。direct-call phrases（119/110）は WAIT48h の例外だが two-step を維持し、自動送信・自動決済・自動通報は使わない。

観測は non-PII KPIs を原則とし、Decision Time／Adherence／revoke 発火率／FPR／FNR を evidence ZIP + SHA256 で保全。weekly SPC でドリフトを監視し、calibration（Brier/ECE）で信頼曲線を整合、decision curve analysis で

operating point の純便益を比較する。更新は「提案→人の承認」で適用し、可逆性を担保。これにより、人は P_res を高めつつ T_eff を短縮し、fatal delay を最小化するための操作余地（作動点・Δt・k/n）を恒常的に保持できる。

SSOT：decision-os-paper / commit SHA= TODO

▼

▶ 7. 家庭マルチシグと図1（BPMN/UML）

案内：図1の主参照は本章。他章は「**§7 図1**」を参照。

【主張の柱（SiriusA Core）】
family multisig を階層型（Cascading）に設計し、Tierの通過で fatal delay を抑えつつ誤介入を最小化する。各Tierは approval window (Δt)、再試行 r、連続ミス閾値 k を持ち、人が operating point を選ぶ。

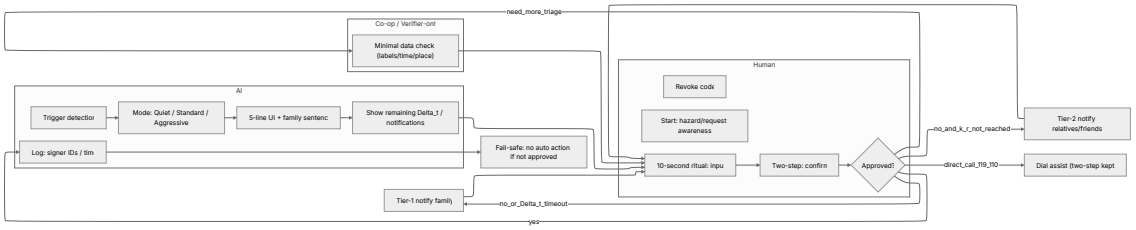

図1：家庭マルチシグ（BPMN/UML風） — 人/AI/Co-opの3レーンと Cascading（Tier-0→Tier-4）。Δt・r・k・revoke を凡例に明記。**未承認＝自動行動なし**、直通語（119/110）は **WAIT48h** 例外でも **two-step** を維持。

【目的】
家族内の合意形成を10秒儀式に接続し、P_res を高めながら T_eff を短縮する安全な承認連鎖を定義する。

1. 階層：Tier-0本人→Tier-1家族→Tier-2親戚/友人→Tier-3 Co-op（Verifier-only）→Tier-4直通語（119/110）。
2. 1/1・1/2対策＝家庭内2-of-2（スマホ＋ウォッチ/据置）。夜間は Δt↑、通知間隔↑、k/r再配分。
3. Co-op は発火時のみ最小データ（ラベル/時刻/場所）を検証。通報権限なし＝倫理境界維持。

a) 図1：BPMN/UMLスイムレーン（人／AI／Co-op）＋Cascading分岐（Δt, r, k）

b) KPI：P_res、FNR、FPR、全Tier貫通失敗率、Decision Time（weekly SPC／calibration=Brier/ECE）

c) ログ：non-PII KPIs と evidence ZIP + SHA256（承認時刻、revocation latency、k/n達成率）

落とし穴（≤2）
・Tierを増やし過ぎると T_eff↑（Quiet既定で抑制）。
・Co-opの監視常時化は信頼崩壊（発火時のみ検証に限定）。

本章は family multisig を階層型に実装する。Tier-0本人の10秒儀式を起点に、未承認のまま approval window (Δt) を越えそうなときのみ上位Tierへ段階的に移行する Cascading を採用する。Tier-1家族は既知の手順と家族一文で P_res を押し上げ、連続ミス閾値 k と再試行 r を満たせない場合に限り Tier-2（親戚/友人）へフォールバックする。Tier-3の Co-op（Verifier-only）は発火時のみ起動し、ラベル/時刻/場所の最小データを検証して誤警報の切り分けを支援するが、通報権限は持たない。Tier-4の direct-call phrases（119/110）は WAIT48h の例外として許容するが、two-step confirmation と revoke code は維持する。

1/1・1/2対策として、家庭内2-of-2（スマホ＋ウォッチ/据置）のデバイス多重署名を基本とし、夜間スロットでは Δt を延長、通知間隔を広げ、k/r を見直すプリセットを適用する。Quiet を既定とし、Standard/Aggressive は人が operating point として明示選択する。観測は non-PII KPIs を原則とし、Decision Time／P_res／FPR／FNR／全Tier貫通失敗率を weekly SPC と calibration（Brier/ECE）で監視、提案されたしきい値変更は人の承認で適用し、evidence ZIP + SHA256 により証跡化する。図1（BPMN/UML）は、人／AI／Co-op のスイムレーンに Δt・r・k・revoke を配置し、Cascadingの分岐条件とフェイルセーフ（未承認＝自動行動なし）を示す。これにより、shared decision-making を保持したまま fatal delay を最小化する。

注：図1凡例に **Δt, r, k, revoke** を明示する（図作成時に反映）。

SSOT：decision-os-paper / commit SHA= TODO

▼ ▶ 8. 安全性検証と運用監査

案内：本章は運用監査（SPC・反証パッチ）に限定。**指標と解析は §9 を参照。**

【主張の柱（SiriusA Core）】

安全は“出荷前”で完結しない。shared decision-making の運用内で、**non-PII KPIs と evidence ZIP + SHA256**を核に、**週次の監視と是正**で継続的に **fatal delay** を下げる。

【目的】

non-PII KPIs を収集→ZIP化→ハッシュ公開し、**weekly SPC 等の監視**と**反証パッチ**で **operating point／approval window (Δt)** を**人の承認**で是正する。

1. **観測**：T_eff／P_res／FPR／FNR／False Alarm Burden（人時）の**最小集合**のみを記録し、設定とログを **evidence ZIP + SHA256** に封緘する。
2. **監視**：週次の統計管理でドリフトを検出し、**較正や純便益比較などの解析は §9 に準拠**して実行する。
3. **是正**：しきい値・UI文言・通知本数などの**最小幅パッチ**を「**提案→人の二重確認→適用**」で反映。AIが自動変更しない。

a) 観測対象：T_eff／P_res／FPR／FNR／False Alarm Burden（人時）

b) **指標・解析：§9 を参照**（KPI定義・calibration・decision curve・SPC は §9 に集約）

c) 追跡：two-step 時刻／revocation latency／family multisig の k/n 達成率（non-PII KPIs＋ZIP+SHA256）

落とし穴（≤2）
・指標の取り過ぎ＝認知負荷とプライバシー負担（**最小集合**を維持）。
・常時監視化＝倫理境界逸脱（Co-op は **Verifier-only／発火時のみ**）。

本章は安全性を「計測→比較→是正」の運用ループとして定義する。観測は **non-PII KPIs** に限定し、**effective time (T_eff)**、**responsiveness (P_res)**、**false negative rate (FNR)**、**False Alarm Burden（人時） を取得する。設定値とログは evidence ZIP + SHA256 に封緘し、改ざん耐性を確保する。監視は weekly SPC 等で行い、遅延伸長・取消過多・応答低下といったドリフトを検出する。解析の詳細（較正／純便益比較／しきい値評価）は §9 に従う。是正は反証パッチ**で実施し、**approval window (Δt)**、通知本数、UI文言・並びを**最小幅**

で修正する。適用は必ず **two-step confirmation と revoke code** を伴う**人の承認**で行い、**自動送信・自動決済・自動通報**は採用しない。**direct-call phrases（119/110）は WAIT48h の例外だが、二重確認は維持する。family multisig は k/n と Δt の実績（達成率、Δt超過率、revocation latency）を継続記録し、提案しきい値は「提案→承認→反映」で反映する。これにより protection of life を保ちながら fatal delay を下げ、逸脱・劣化・虚偽を再現可能ログ**で追跡できる。

— **8.1 非個情KPIの収集・ZIP化・ハッシュ公開**：KPI最小集合を収集し、**evidence ZIP + SHA256** を公開可能形で保全（**定義・計算は §9 準拠**）。

— **8.2 週次監視と反証パッチ**：**weekly SPC 等**でドリフト検出→**人の承認**を経てパッチ適用（**自動変更なし**）。

— **8.3 監査可能性（再計算）**：two-step／revoke／k/n／Δt／作動点を揃え、**外部再計算**で一致を保証（**解析手順は §9 参照**）。

SSOT：decision-os-paper / commit SHA= TODO

▼ ▶ 9. 評価計画（KPI・プロトコル）

正典：KPIの名称・定義・取得手順、解析手法（calibration, decision curve, SPC）は本章のみで定義する。

KPI

1. **Adherence**：規約どおり手順が守られた比率（%）。
2. **Decision Time**：承認完了までの秒数。
3. **P_res（反応性）**：提示→操作までの応答性（高いほど速い）。
4. **FPR（誤警報率）**：偽陽性の割合。
5. **FNR（見逃し率）**：偽陰性の割合。
6. **False Alarm Burden（人時）**：誤警報対応に要した総人時間。
7. **Net Benefit（decision curve）**：しきい値確率に対する純便益。
8. **calibration（Brier/ECE）**：予測と実測の整合。

【主張の柱（SiriusA Core）】
評価は「ケース×作動点×時間帯」で設計し、non-PII KPIs と証跡（evidence ZIP + SHA256）に基づく再現可能な比較で、protection of life に直結する fatal delay・FPR/FNR・Net Benefit を同時に検証する。

【目的】
ケース別の expected harm minimization を検証し、operating point

（Quiet/Standard/Aggressive）と approval window (Δt) の推奨レンジを提示する。

1. ケース設計：詐欺／災害／転倒／家族連絡を各3トライアル（計12）。時間帯は〔深夜／昼〕の2スロットで分散、order effect はラテン方格で制御。
2. KPI取得：Adherence、Decision Time（秒）、P_res、FPR、FNR、False Alarm Burden（人時）、Net Benefit、calibration（Brier/ECE）。
3. 解析：decision curve analysis で作動点の純便益を比較し、weekly SPC でドリフト監視。提案しきい値は人の承認で更新。

a) 設計：ケース×作動点×時間帯の要因計画／Δt・k/n・通知本数の前登録

b) 指標：Adherence／Decision Time／FPR／FNR／P_res／Net Benefit／Brier／ECE

c) 証跡：非個情KPIログ＋設定を evidence ZIP + SHA256（前登録ID・コミットSHA付き）

落とし穴（≤2）
・Quiet/Standard/Aggressive の順番固定＝学習バイアス（ラテン化必須）。
・revocation latency の未計測＝P_res 過大評価（常時記録）。

評価は、ドメイン4種〔詐欺／災害／転倒／家族連絡〕×各3トライアルの計12タスクで行う。各トライアルは operating point（Quiet/Standard/Aggressive）をラテン方格で割り付け、order effect を抑制する。時間帯スロットは〔深夜／昼〕で均等化し、approval window (Δt)・family multisig（k/n, Δt）・通知本数は事前登録（pre-registration）で固定する。入力は「10秒儀式（5行UI）」に限定し、two-step confirmation と revoke code を必ず通す。

取得する non-PII KPIs は Adherence、Decision Time（承認完了までの秒）、P_res、FPR、FNR、False Alarm Burden（人時）、Net Benefit、calibration（Brier/ECE）。直通語（119/110）は WAIT48h の例外だが two-step を維持し、自動送信・自動決済・自動通報は採用しない。欠測は Δt超過として扱い、未承認＝自動行動なしを厳守する。revocation latency は revoke code 入力から無効化完了までの秒で測定する。

解析は decision curve analysis を主軸に、閾値確率レンジにわたる Net Benefit の差を作動点間で比較し、ケース別の推奨 operating point と Δt を導出する。weekly SPC では時系列のドリフト（T_eff 伸長、P_res 低下、FPR/FNRの偏り）を監視し、calibration（Brier/ECE）で予測と実測の整合を評価、必要な反証パッチ（Δt・通知本数・UI文言）を生成する。ただし適用は「提案→人の承認」で実施し、shared decision-making を保持する。

すべてのログと設定は evidence ZIP + SHA256 に封緘し、前登録IDとSSOT（GitHubのSHA）に紐づけて公開可能形に整える。アウトカムは「10秒内達成率」「全Tier貫通失敗率」「Net Benefit 優越」 を主要指標として提示し、expected harm minimization に資する実用的な作動点レンジを結論として返す。

SSOT：decision-os-paper / commit SHA= TODO

▼ ▶ 10. 関連研究

【主張の柱（SiriusA Core）】
時間制約下のHRI/HF、AI倫理のHuman-in-the-loop、Gerontechnologyの可用性研究を横断し、SiriusAは「10-second ritual」と shared decision-making を中核に、protection of life を実装規約として具現化する。差分は、operating point（Quiet/Standard/Aggressive）と approval window (Δt) を“人が選ぶ”ことで fatal delay を実運用で最小化する点。

【目的】
既存知を整理し、SiriusAの新規性＝時間最適化（T_eff/P_res）と倫理境界の両立を明確化する。

1. HRI/HF：警告設計・二段確認・触覚/視覚/音の多様提示はあるが、T_eff と P_res を指標連関で最適化した枠組みは希薄。
2. AI倫理：Human-in-the-loopは概念中心。SiriusAは two-step confirmation ＋revoke code を義務化し、auto行為を排した運用規約に落とす。
3. Gerontechnology：高齢者の認知負荷研究は豊富だが、family multisig と 10-second ritual を接続した“家族合意×時間最適化”は未整備。

a) 指標連関：T_eff＝Δt_set＋P_res^-1、FPR/FNR、Net Benefit

b) 手法：decision curve analysis／calibration（Brier/ECE）／weekly SPC

c) 記録：non-PII KPIs と evidence ZIP + SHA256（再計算可能性）

落とし穴（≤2）
・関連研究の“理想UI”前提を鵜呑みにすると現場乖離（運用規約が必要）。
・倫理議論のみで時間最適化を欠くと fatal delay が残存。

時間制約下のHRI/HFは、注意喚起・段階的確認・モダリティ統合の有効性を示す一方、operationalな最適化（T_eff と P_res の同時最小化）を中心に据えた設計は限定的である。SiriusAは approval window (Δt) と operating point（Quiet/Standard/Aggressive）を“人が選ぶ”構造に固定し、fatal delay を expected harm minimization の観点で扱う。AI倫理のHuman-in-the-loopは抽象的原則として広がったが、SiriusAは two-step confirmation と revoke code を強制要件とし、direct-call phrases（119/110）も WAIT48h 例外としながら二重確認を保持する運用規約で具体化する。Gerontechnologyは高齢者の認知負荷・可用性を詳細に報告してきたが、family multisig により“家族の到達可能性”を制度化し、10-second ritual によって意思確定の可逆性を担保する統合は乏しい。本研究は non-PII KPIs を evidence ZIP + SHA256 で保全し、weekly SPC と calibration（Brier/ECE）、decision curve analysis によって T_eff／P_res／FPR／FNR／Net Benefit を再現可能に評価する点で、既存知を“規約＋計測”に架橋する。

SSOT：decision-os-paper / commit SHA= TODO

▼ ▶ 11. 限界・リスク・法務

案内：ゼロリスクを前提にしない。残余 **FPR/FNR** と **T_eff/P_res** を併記し、選択は人が行う。

【主張の柱（SiriusA Core）】
100%安全は存在しない。残余の false positive rate (FPR)／false negative rate (FNR) と期待被害を可視化し、shared decision-making を守りつつ protection of life を最大化する。

【目的】
non-medical の限界を宣言し、責任分界と証跡で説明責任を果たしつつ、expected harm minimization を運用規約に落とす。

1. 残余リスク管理：FPR/FNR と effective time (T_eff)／responsiveness (P_res) を併記し、operating point と approval window (Δt) を人が選ぶ。

2. 法的境界：医療判断は外部（119／医師等）へ委譲。自動送信・自動決済・自動通報は採用しない。
3. 説明責任：non-PII KPIs と evidence ZIP + SHA256 により、意思決定と取消の再現可能性を担保。

a) 指標：FPR／FNR／T_eff／P_res／False Alarm Burden（人時）

b) 手法：weekly SPC／calibration（Brier/ECE）／decision curve analysis

c) 記録：two-step 時刻／revocation latency／family multisig（k/n, Δt）ログ

落とし穴（≤2）
・“ゼロリスク”を暗黙前提にすると運用が崩壊（閾値と作動点の公開が必須）。
・常時監視化は法倫理逸脱（Co-opは Verifier-only、発火時のみ）。

本章は non-medical としての限界を明示する。SiriusA は fatal delay を下げるが、FPR と FNR をゼロにはできない。したがって、残余リスクを T_eff／P_res と併記し、operating point（Quiet/Standard/Aggressive）と approval window (Δt) を“人が選ぶ”前提を維持する。医療判断は direct-call phrases（119/110）や医師へ委譲し、WAIT48h 例外であっても two-step confirmation と revoke code を保持する。自動送信・自動決済・自動通報は採用しない。

責任分界は「提案＝AI／決定＝人」で固定する。取消（revoke）と二重確認の双方時刻、family multisig（k/n, Δt）、承認者Tier、作動点の記録を non-PII KPIs として収集し、evidence ZIP + SHA256 に封緘する。weekly SPC でドリフトを検出し、calibration（Brier/ECE）で較正、decision curve analysis により作動点の純便益を比較する。改善は「提案→人の承認」で適用し、再計算可能性により説明責任を果たす。社会実装では、家族・企業・Co-op・自治体の利害調整を前提とし、監視の常態化を避けるため Co-op は Verifier-only／発火時のみとする。以上により、期待被害を最小化しつつ、人の尊厳と法倫理の境界を維持する。

SSOT：decision-os-paper / commit SHA= TODO

▼ ▶ 12. 結論と今後（PoC／標準化）

【主張の柱（SiriusA Core）】
目的は protection of life。SiriusA は 10-second ritual と shared decision-

making を核に、fatal delay を T_eff／P_res の最適化で下げる。次は PoC と標準化で運用規約を社会実装へ接続する。

【目的】
D0→D7→D30→D90 の段階計画で仮説を検証し、閾値秘匿・証跡公開を原則に標準化を前進させる。

1. ロードマップ：D0（本稿）→D7（小規模PoC準備）→D30（PoC実施）→D90（拡張・監査）。
2. PoC：operating point（Quiet/Standard/Aggressive）と approval window (Δt) の家庭別最適化仮説を検証。
3. 標準化：閾値は秘匿、non-PII KPIs と evidence ZIP + SHA256 は公開。SSOT（GitHubのSHA）で一意化。

a) KPI：T_eff／P_res／FPR／FNR／Net Benefit／calibration（Brier/ECE）

b) 手法：decision curve analysis／weekly SPC（ドリフト検知）

c) 運用：two-step／revoke／family multisig（k/n, Δt）ログの前登録と再計算

落とし穴（≤2）
・自動適応の過信＝人の承認喪失。
・公開指標の過多＝プライバシー漏れ（最小集合維持）。

本稿（D0）は、T_eff＝Δt_set＋P_res^-1 を軸に、operating point と approval window (Δt) を“人が選ぶ”設計で fatal delay を最小化する枠組みを示した。今後は D7 で PoC計画を確定し、ケース（詐欺／災害／転倒／家族連絡）×作動点の小規模検証を整える。D30 で PoC を実施し、non-PII KPIs（Decision Time／P_res／FPR／FNR／Net Benefit／Brier／ECE）を収集、weekly SPC と decision curve analysis で推奨作動点と Δt を導出する。D90 では拡張と監査体制を確立し、閾値は秘匿、evidence ZIP + SHA256 と SSOT（GitHubのSHA）で証跡公開の原則を固定する。研究課題は、作動点適応の自動化を**人の承認**（two-step＋revoke code）で必ずゲートする方式の最適化、family multisig の k/n と時間帯プリセットの一般化、False Alarm Burden（人時）を含む期待被害最小化の実地妥当性である。SiriusA は標準化（用語・計測・証跡）を先行し、社会実装へ段階的に接続する。

SSOT：decision-os-paper / commit SHA= TODO

▼ ▶13.最終章

1. Architectural Ethics（設計としての倫理）

本研究は、責任を固定的な帰属ではなく**時間的に移動するベクトル**として扱う。SiriusAは**二重確認（two-step confirmation）・取消コード（revoke code）・WAIT48hの例外維持**を備え、致死的遅延（fatal delay）を10秒儀式（10-second ritual）で抑制する運用倫理を実装する。作動点（operating point：Quiet/Standard/Aggressive）は人が主権的に選択し、承認窓（Δt）**を管理する。これにより「誰が決めたか」に加え、「いつ・どの手順で決まったか」が**証跡ZIP＋SHA256（evidence ZIP + SHA256）で監査可能となる。

2. Multi-Model Convergence Principle（多モデル収束原理）

複雑仮説の信頼性は単一モデルの尤度ではなく、**異設計AIの収束（合意）で評価する。実務ではagreement ≥ τ**（例：**Cohen's κ、平均JSD**）を**事前登録**し、未達時は**作動点をQuiet側へ一段**倒して**二重確認を強化**する。モデル名・版・プロンプトは付録に記録し、**最終承認は人**が行う。これにより**検証容易性**と**再現性**を両立する。

3. Self-Amplifying Structure（時間が味方する理論）

AIの**性能と多様性**が増すほど、収束検証の**探索空間は拡大**し、理論は**時間とともに単調強化**される。分散システムの信頼蓄積に類比可能で、**単一障害点**への依存を避ける。週次SPC（weekly SPC）**と**較正（calibration：Brier/ECE）により、反応性（P_res）と安全余裕のバランスを運用で最適化する。

4. Synthesis and Implications（統合と含意）

SiriusAは**時間拘束下の運用倫理・多モデル収束・時間強化性**を束ね、**"10秒での安全な行動"を目的関数とするパラダイムを提示した。家庭マルチシグ（family multisig：k/n, Δt）と非個情KPI（non-PII KPIs）を核に、方法・結果を改変せず社会実装可能な規約へ落とし込む。今後は領域横断PoCで閾値τと人の反応性分布**を実地較正し、**標準化：閾値は秘匿・証跡は公開**の原則へ接続する。
〔SSOT: decision-os-paper/**<commit-SHA>**〕

SSOT：decision-os-paper / commit SHA=TODO