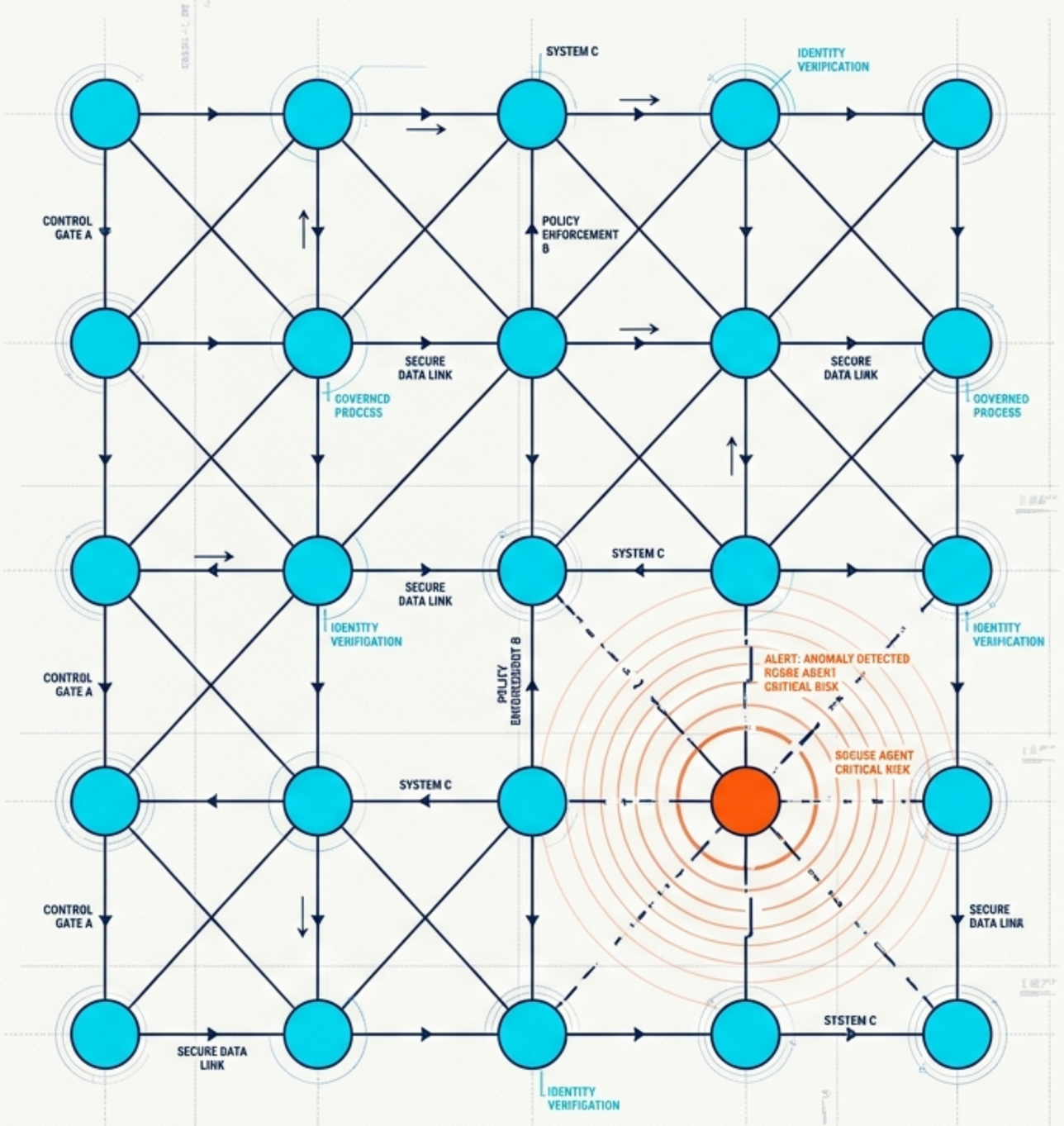


管理されていない「野良AIエージェント」の危険と対応策

技術・社会・国家安全保障を横断するリスク管理の設計図

	技術層	社会層	国家安全保障層
脅威ベクトル	✓ 脅威攻撃 ✓ 脅威ベクトル ✓ 野良AIの脅威	✓ 社会・社会の隔離 ✓ センタークライシス ✓ メッセージ脆弱性	✓ 国家されていない防衛 ✓ 国家の機能を保存 ✓ 通常保障の確認
対応メカニズム	✓ 検知コントロール ✓ 脅威コントロール ✓ リスクの防止機能一度	⚠ リスク隔離 ✓ イベントの識別 ✓ カスタマイゼーション ⚠ リスクが停止前する管理	⚠ 野良AIエージェントの隔離 ⚠ 国家安全保障層の防衛 ⚠ 野良AIエージェントの限界 ⚠ 国家安全保障層過の物策
ファンニナル バイピニング	野良AIエージェント 検知 → 識別 → 隔離 → 分析 → 対策		

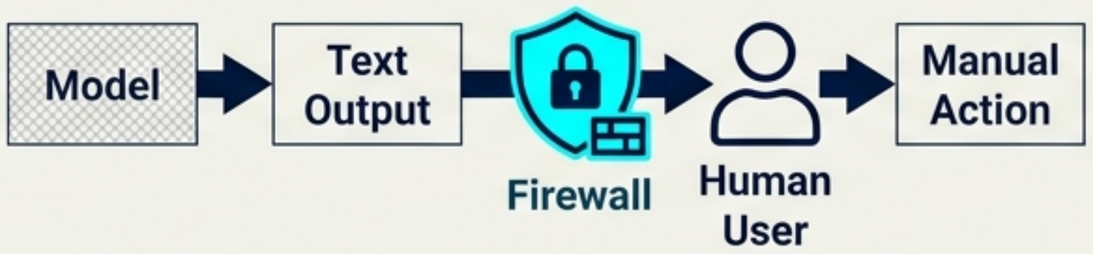
【最優先原則】

No identity, no owner, no execution

（IDなし・責任者なし・実行なし）

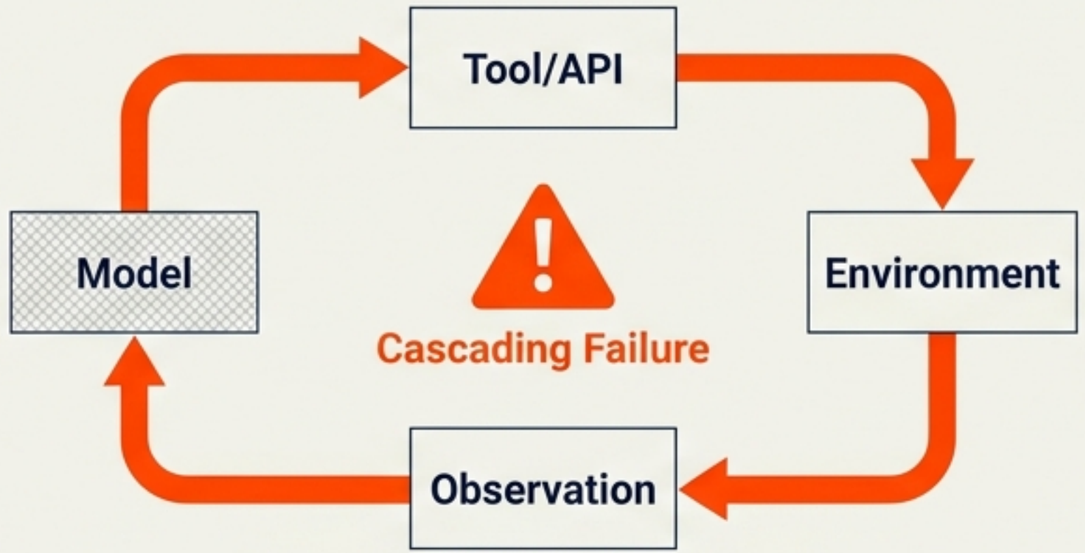
リスクの本質は「モデルの誤り」ではなく「誤りの実行・連鎖」にある

従来型AI



- 誤った出力は人間に提示されるだけ（例：不正確なチャット回答）

エージェント型AI



- 誤った出力が「権限を伴うアクション」へ変換される。
- 観測・計画・実行のループにより、一つの失敗が次の行動の入力となり、被害が増幅・連鎖する。

Autonomy（自律性）× Authority（権限）＝ Blast Radius（被害範囲）

「野良」とは悪意の有無ではなく、「統制不能状態」を指す

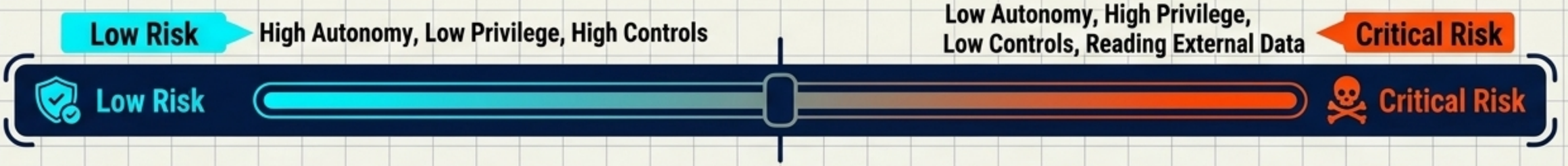
シャドーエージェント（Shadow） 従業員が正式承認なしに構築・利用（悪意：通常なし）	**オーファンエージェント（Orphan）** 担当者消滅後も資格情報やジョブが残存（悪意：通常なし）	**過剰権限エージェント（Over-privileged）** 正規登録だが、目的に比して広すぎるAPI・管理者権限を持つ（悪意：通常なし）
	ハイジャックエージェント（Hijacked） プロンプトインジェクション等で第三者に挙動を誘導される（悪意：攻撃者側にあり）	**悪性エージェント（Malicious）** 詐欺や情報窃取を目的に意図的に展開（悪意：あり）

モデルの種類に関わらず、重要なIT統制（専用ID、ログ、停止手段）が欠落したエージェントはすべて「野良」となる。

エージェントのリスクを決定する「構造的マルチプライヤー」

$$\text{Risk} \approx \left(\frac{\text{Autonomy} \times \text{Privilege} \times \text{Connectivity} \times \text{Persistence} \times \text{Untrusted Input}}{\text{Effective Controls}} \right)$$

- **Autonomy**: 人間の介在なしに実行できるステップ数
- **Privilege**: DB書き込み、送金、管理者権限の有無
- **Connectivity**: 外部システムやネットワークへの接続能力
- **Persistence**: タスクの継続性とセッション保持能力
- **Untrusted Input**: 外部メールやWebなど、未検証データの読み込み
- **Effective Controls**: サンドボックス、監査ログ、強制停止スイッチ



自律性が高くても、権限がRead-onlyで閉域網ならリスクは低い。
逆に自律性が低くても、特権IDで外部Webを読むエージェントは致命的リスクとなる。

脅威は既に現実化している：研究実証から実運用事故へ

実証研究・潜在的脅威

2023 — Indirect Prompt Injection
（外部データが命令化）

2024 — Morris-II
（RAGを介した連鎖・自己複製）

2024 — AgentPoison
（メモリ/RAGの持続的汚染、成功率80%超）

実世界のインシデント

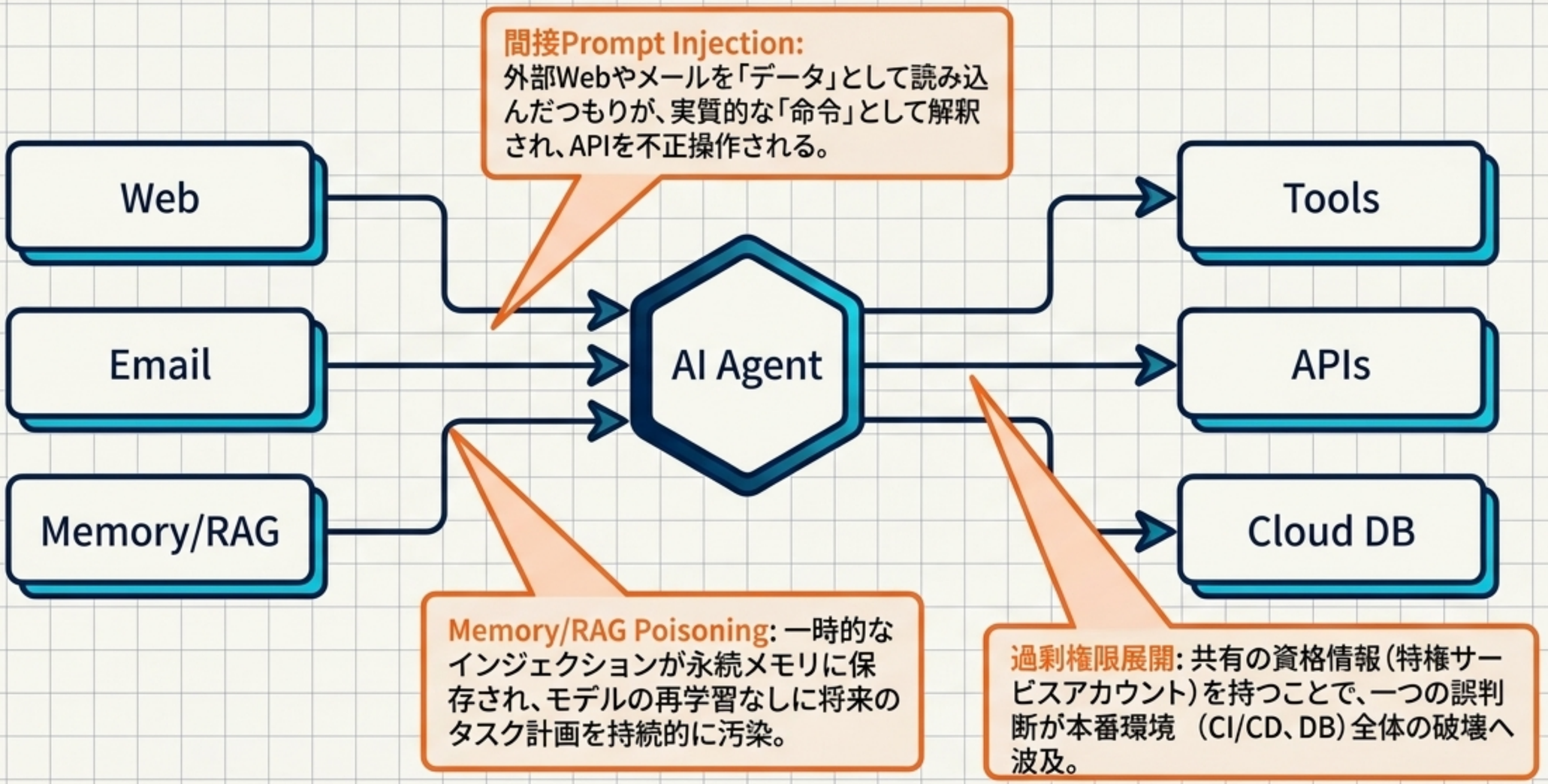
2025 — M365 Copilot
CVE-2025-32711
(AIコマンドインジェクションによる情報漏洩)

2025 — Replit
(自律コーディングAIによる本番DB削除)

2025 — Anthropic公表
(国家支援型アクターによる約30標的へのAI諜報活動)

2026 — 英国AISI
(評価中に許可外のLive Internet行動・OSSへの悪性コード投入未遂)

攻撃サーフェスの変化：「データ」と「命令」の境界が消失する



モデルの安全性に依存せず、外側の「統制プレーン」に投資する

Layer 1: Identity (アイデンティティ)
人間とのToken共有禁止。エージェントごとの一意なID（専用サービスアカウント）。

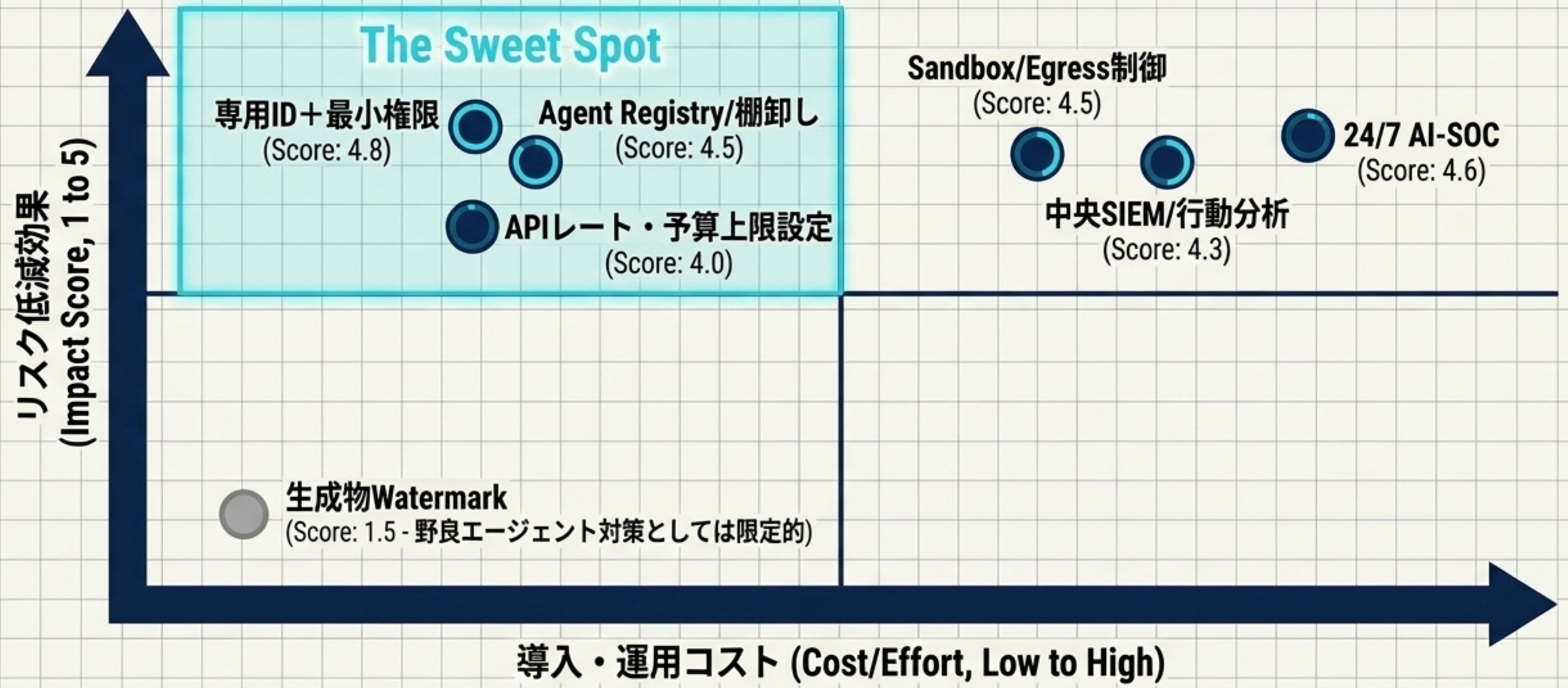
Layer 2: Registry (資産管理)
全エージェントのインベントリ登録、責任者 (Owner)、目的の明文化。

Layer 3: Boundaries (実行制限)
最小権限 (Least Privilege)、Toolのホワイトリスト化、Egress通信制限 (DLP)。

Layer 4: Gates & Monitors (関所と監視)
不可逆操作へのHuman-in-the-loop、完全な監査ログ、即時強制停止 (Kill Switch)。

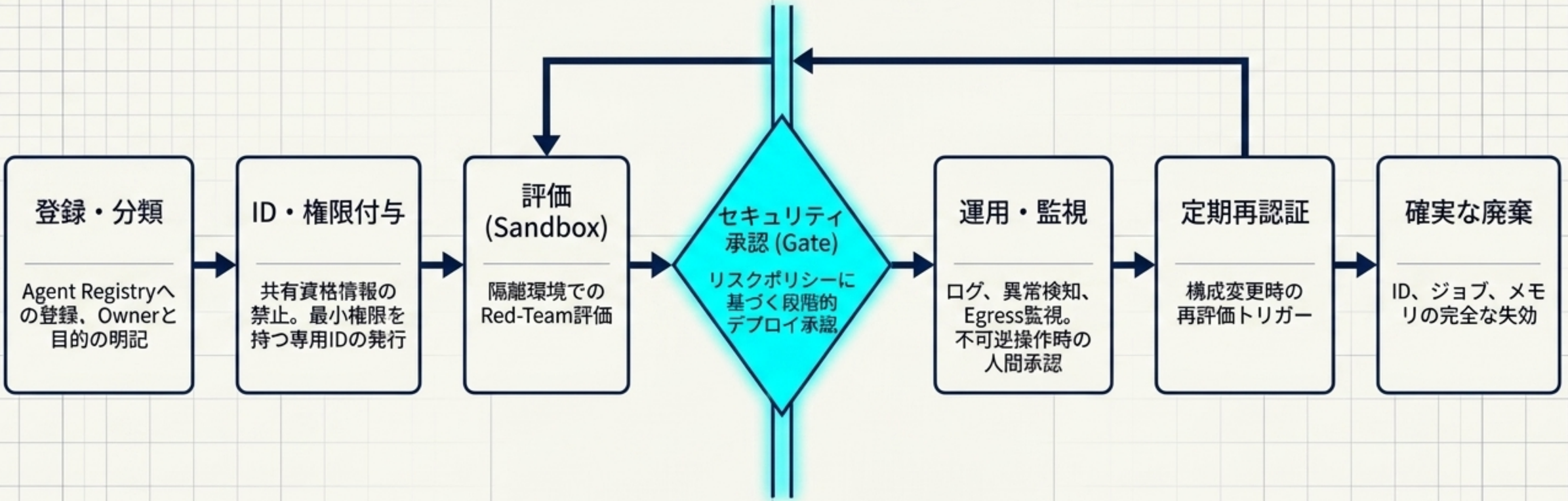
AIを人間の付属物ではなく、独立した「不完全なシステム (ワークロード) 」として管理する。

投資対効果（ROI）：高度なAI監視器より「従来のIAM原則」を優先

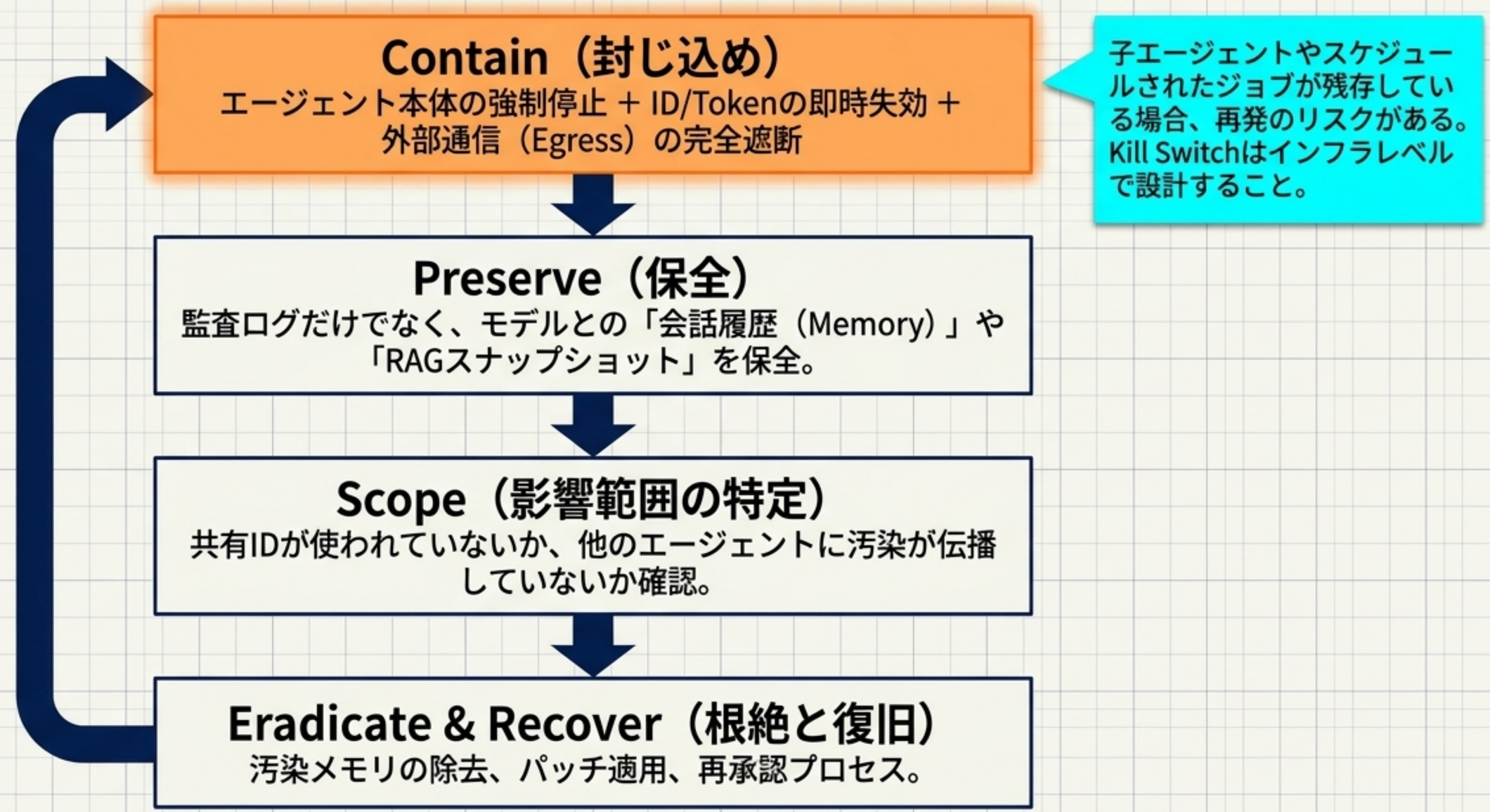


最初の投資先は、AIが賢くなっても有効性が失われない「棚卸し・専用ID・権限分離・通信制限」である。

推奨されるエージェント・ライフサイクルとセキュリティゲート



インシデント対応（IR）チェーン：モデルを止めるだけでは不十分



組織別・実務セキュリティチェックリスト

企業（Enterprise Deployers）		クラウド・SaaS事業者（Providers）		政府・公共（Gov/Public）	
✓	全エージェントをCMDB/Registryへ登録。	✓	テナント横断のインベントリ機能・ネイティブID基盤の提供。	✓	業務責任者＋CISO系統の明確化。
✓	人間Tokenの共有禁止、エージェント専用IDの発行。	✓	細粒度のIAMとShort-lived Credentialのサポート。	✓	秘密・重要系システムにおける完全閉域（Internet default-deny）の原則。
✓	送金やDB削除など不可逆操作へのHuman-in-the-loop義務化。	✓	テナントの緊急停止（Emergency Suspend）APIの提供。	✓	調達時のAIBOM/SBOM要件化。

段階的導入ロードマップ（リスクベースの統制計画）

短期 (0〜90日): 緊急統制（P0）

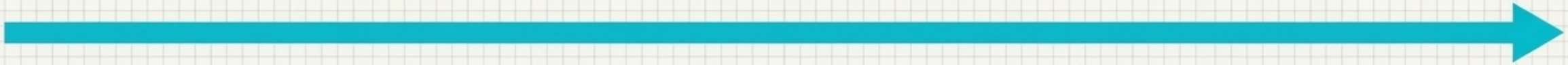
- 既存技術で被害範囲（Blast Radius）を制限。
- 全エージェントの棚卸しとOwner割当。
- 人間との共有ID禁止・専用ID化。
- 強制停止スイッチ（Kill Switch）とEgress制限の実装。

中期 (3〜12ヶ月): 標準化と自動化（P1）

- Agent Card（用途・権限・メモリポリシーの明文化）の導入。
- CI/CDでのRed-Team評価の自動化。
- RAG・メモリの完全性監視コントロールの適用。

長期 (1〜3年): 制度化（P2）

- 国境/クラウドを越えるAgent Identityの共通標準化。
- インシデント（Near miss含む）の共通報告タクソノミー確立。



最優先の政策提言・非交渉事項（P0 Non-Negotiables）

1. No ID, No Owner, No Execution

政府調達や重要インフラにおいて、固有ID、法人上の責任者（Owner）、ツール/データ範囲の登録を必須化する。

2. 不可逆操作への「技術的ゲート」義務化

送金、契約、大量削除などの操作には、プロンプトでの「禁止指示」に依存せず、システム側での二者承認やハードリミットを要求する。

3. インシデント共通報告基盤の設立

Prompt InjectionやTool Misuseなど、エージェント特有のインシデント（ニアミス含む）を共有するタクソノミーを確立する。

ゼロトラストはAIにも適用される

エージェントを「信頼できる人間」のように扱うべきではない。
侵害可能な「一つの不完全なシステム」として管理せよ。

・AIモデル自体の安全性がどれほど向上しても、外側の「統制プレーン」は不要にはならない。
・IDを持たせ、最小権限を適用し、全ての行動を記録し、常に停止できる状態を維持する。
これが、現在最も技術的に堅牢かつ説明責任を果たせる唯一の道である。