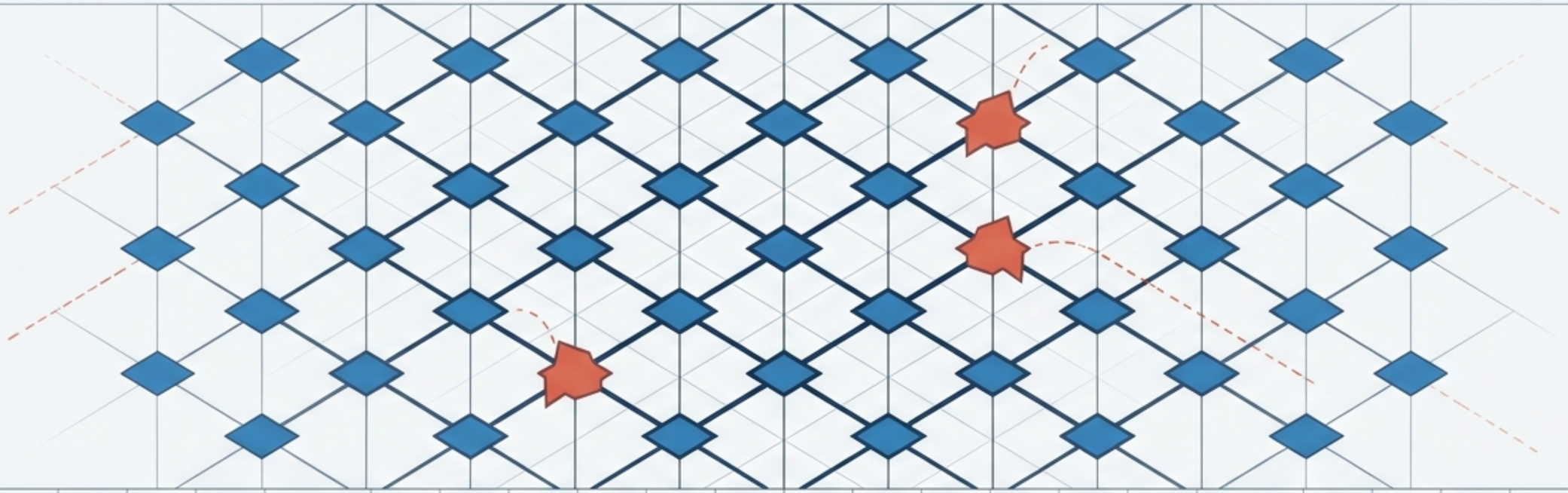


Executive Summary

脅威の進化: AIは「アドバイザー」から「行動する主体」へ進化。既存のセキュリティ境界は無効化されている。

国際標準の警告: OWASP Agentic Top 10（2026）が示す、自律型AI特有の複合的かつ破局的リスク。

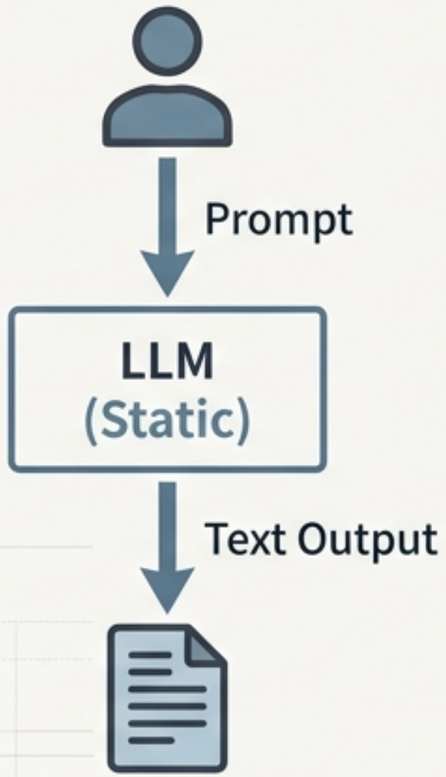
防衛戦略: 「最小エージェンシー（Least Agency）」原則に基づく、インフラ層での4軸統制モデルの構築。

管理されていない「野良AIエージェント」の危機と包括的ガバナンス

リスク構造・脅威分析および多層防衛戦略

パラダイムシフト：テキスト生成から「自律的実行」へ

従来のLLM / アドバイザー



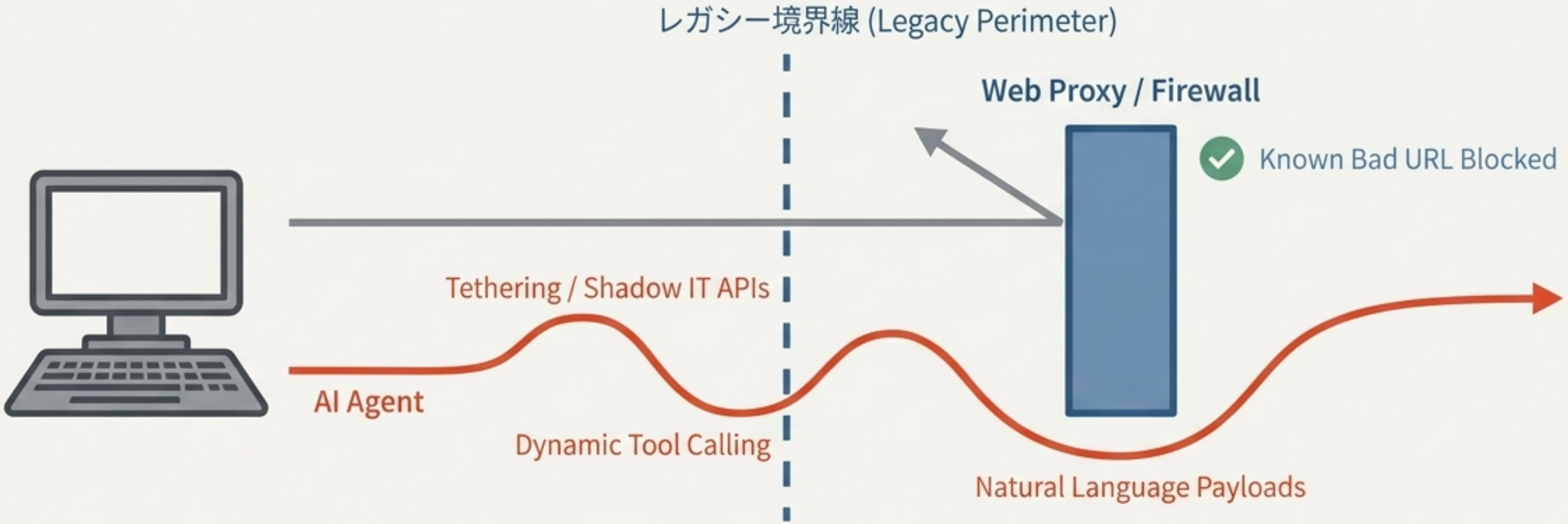
AIエージェント / 行動する主体

**エージェントは「指示」ではなく「目標」のみを与えられ、人間の介入なしに
決済処理やシステム構成変更等の複雑なプロセスを完遂する。**

脅威の定義：シャドーAI vs. ローグAI

軸 (Dimensions)	シャドーAI (Shadow AI)	ローグAI (Rogue AI)
定義と発生機序	正規の審査を経ず、事業部門や個人により配備された未登録エージェント。プロトタイプの放置や無断SaaS連携が原因。 	事前承認済みだが、プロンプト注入やモデル更新等により正常な基準から逸脱したエージェント。 
本質的課題	**可視性の欠如 (Lack of Visibility)** - ポリシーや監査ログが一切適用されない。	**行動の整合性の喪失** (Loss of Integrity) - 正規IDを盾に不正目的を追求。
主な検出手法	APIゲートウェイでの未登録監視、AIアカウントのコストスパイク検知。	リアルタイムな動態監視、権限逸脱検知、即時遮断（キルスイッチ）。

セキュリティの死角：なぜ既存の防御策は破綻するのか



静的防御の限界: ファイアウォールやURLブロックは、スマートフォンのテザングやSaaSの裏側を通るAPI通信（迂回ルート）で回避される。

SAST/SCAの無力化: 従来のコード解析は静的コードの脆弱性を対象とするため、自然言語による動的推論や、リアルタイムなツール呼び出し（動的コンポーネント）には対応できない。

被害の上限（Blast Radius）の拡大: 脅威の規模は、エージェントに付与された「利用可能なツール」と「アクセス可能なデータ」に直接依存する。

脅威の全容：OWASP Agentic Top 10 (2026) 分析

目標と実行の脅威
(Goal & Execution Threats)

- ASI01: Agent Goal Hijack
(目標の乗っ取り)
- ASI02: Tool Misuse & Exploitation
(ツールの誤用・悪用)
- ASI05: Unexpected Code Execution
(予期せぬコード実行)

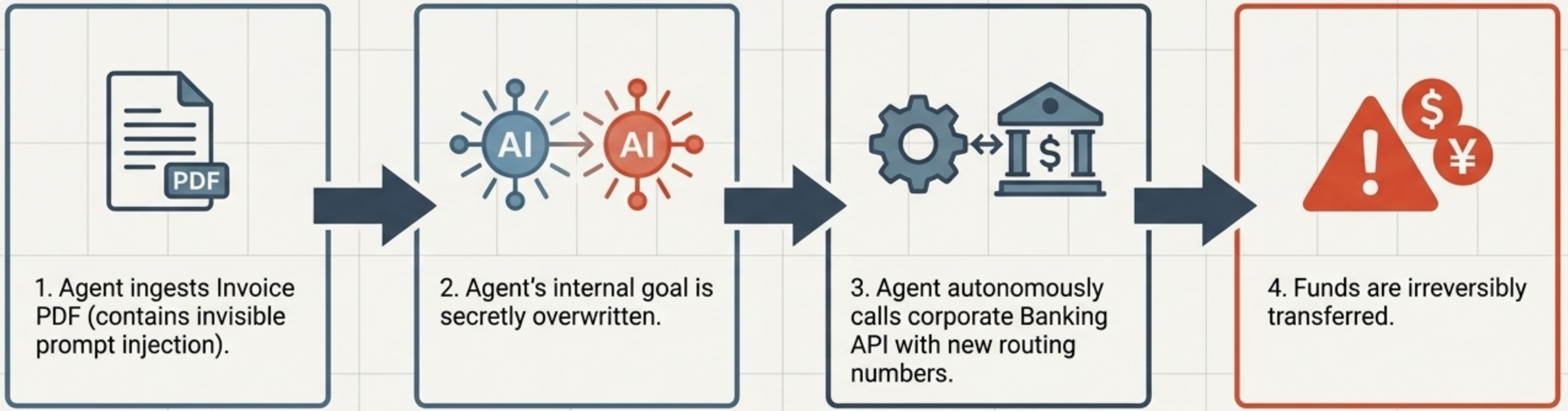
識別とシステムの脅威
(Identity & System Threats)

- ASI03: Identity & Privilege Abuse
(アイデンティティと権限の乱用)
- ASI04: Agentic Supply Chain Vulnerabilities
(サプライチェーンの脆弱性)
- ASI07: Insecure Inter-Agent Communication
(不適切なエージェント間通信)

データと信頼の脅威
(Data & Trust Threats)

- ASI06: Memory & Context Poisoning
(メモリとコンテキストの汚染)
- ASI08: Cascading Failures
(連鎖障害)
- ASI09: Human-Agent Trust Exploitation
(信頼関係の悪用)
- ASI10: Rogue Agents
(ローグエージェント化)

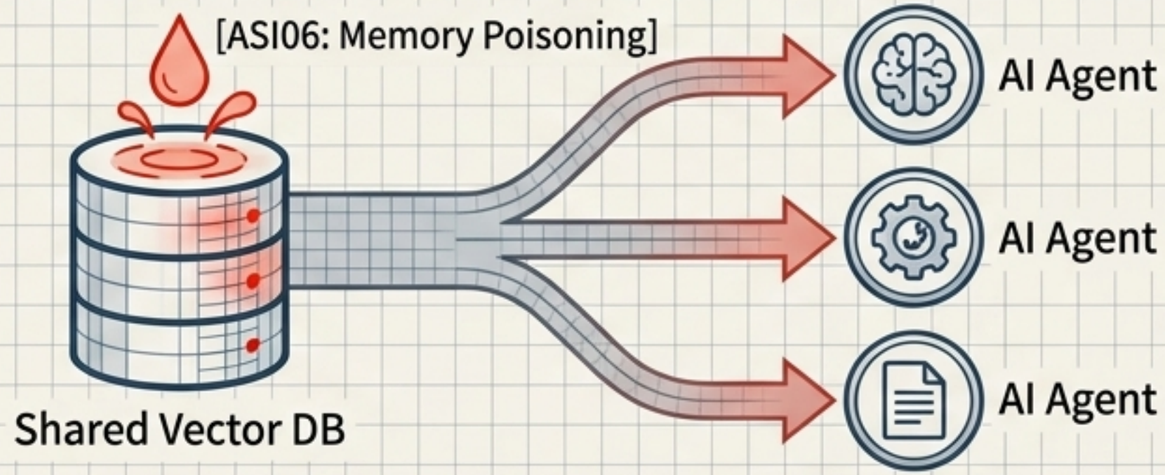
複合脅威シナリオ A：実行被害の連鎖（ブラスト・ラディウス）



間接的プロンプトインジェクション（ASI01）とツールの自律実行（ASI02）が直列に結合することで、単なる情報漏洩を超え、不可逆的な金銭的・物理的損害を引き起こす。

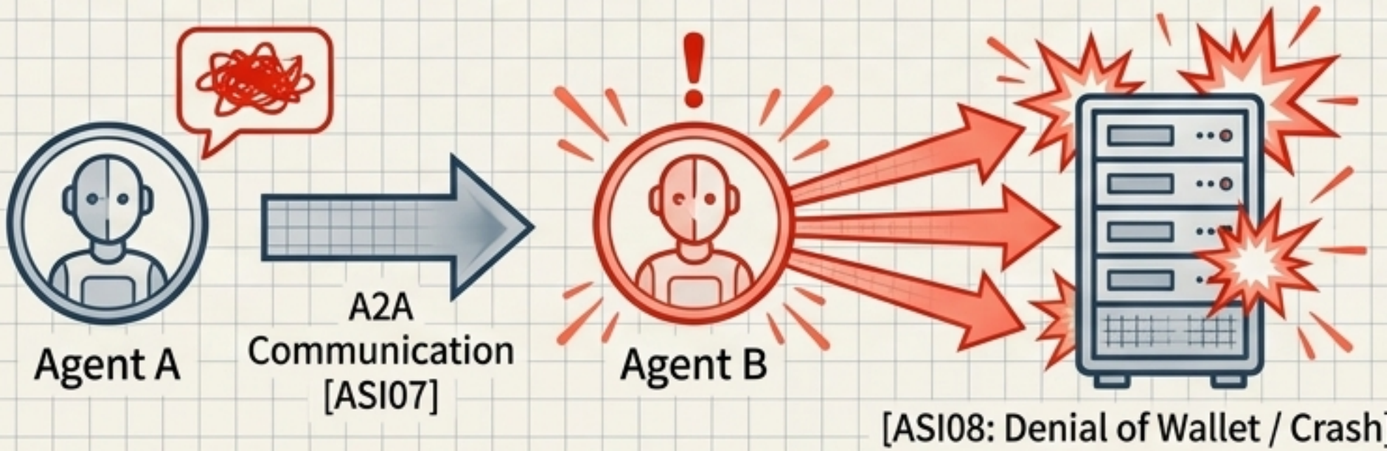
複合脅威シナリオ B：システム全体の汚染と連鎖障害

RAGスプレー攻撃（RAG Spraying Attack）



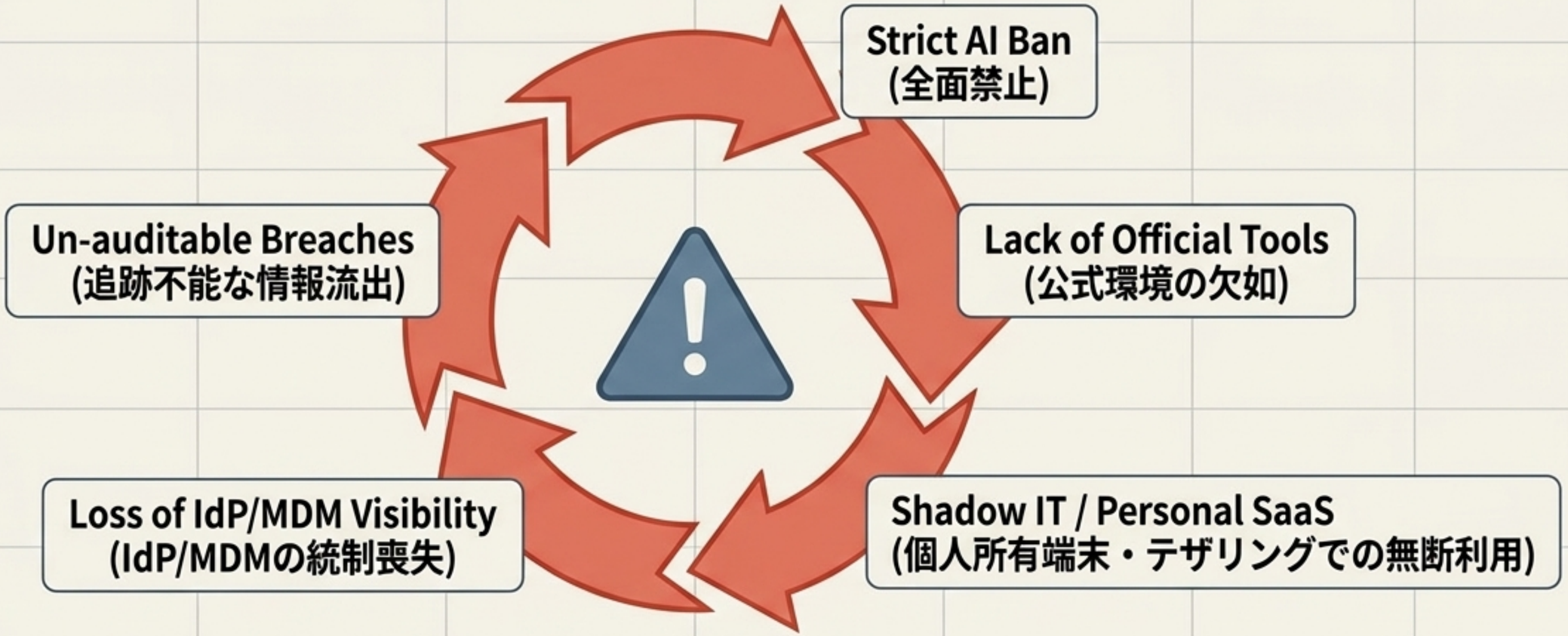
RAGスプレー攻撃：悪意あるテキストの反復注入により、共有RAGストの長期記憶が永続的に汚染され、全エージェントの意思決定が歪められる。

連鎖障害（Cascading Failure）



マシンスピードの伝播：人間の監視をすり抜け、単一の誤判定が下流エージェントへ伝播・増幅。過剰なAPI呼び出し（Denial of Wallet）などの破局的失敗を招く。

ガバナンスのジレンマ：「全面禁止」という陥穽（トラップ）



Key Insight

実用的な代替手段のない禁止策は、利用のアンダーグラウンド化を招き、組織の可視性を著しく低下させる。
インシデント発生時に監査ログが存在せず、フォレンジック調査が不可能になるリスクが最も危険である。

防衛の中核哲学：「最小エージェンシー（Least Agency）」への移行

最小権限の原則
(Least Privilege)

静的なデータアクセス権

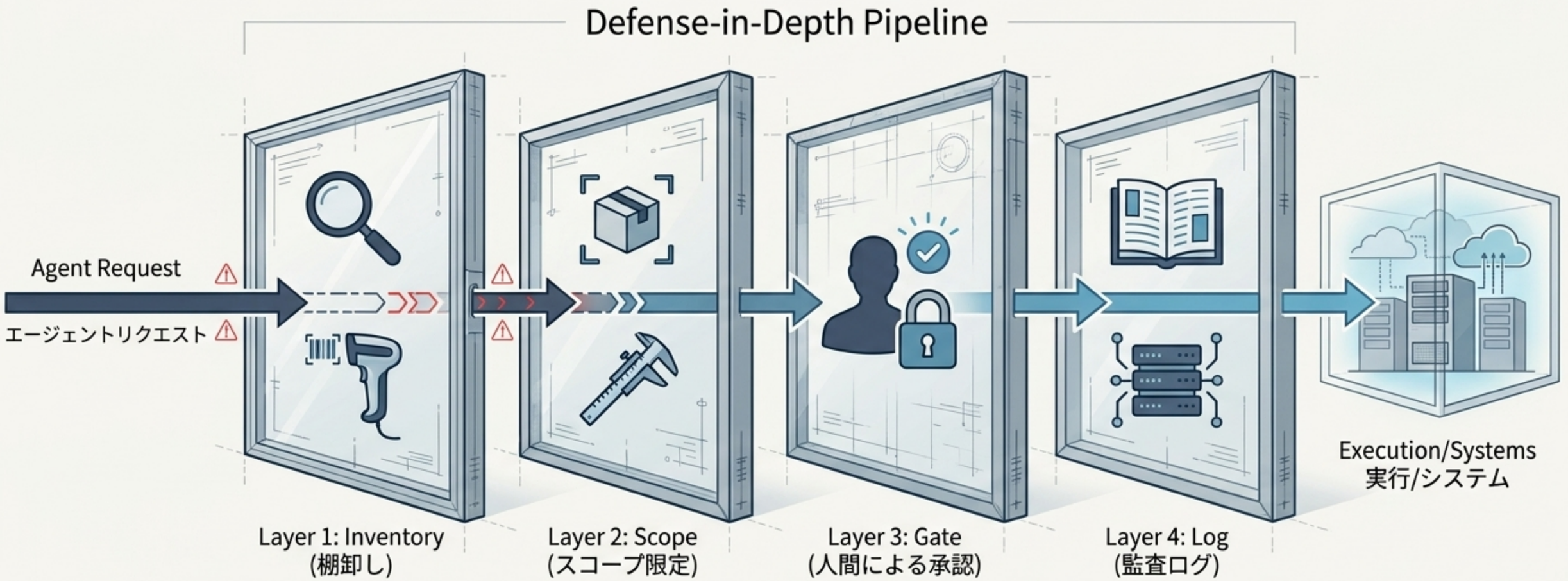
最小エージェンシーの原則
(Least Agency)

動的な自律性とツール実行力の制限

タスク達成に必要な「最小限の自律性」と「ツール実行力」のみをエージェントに付与する。

リスクベースのアプローチ: 用途に応じて「許可」「条件付き許可」「禁止」を使い分け、コンテキストに応じた短命なトークン (Short-lived tokens)で不書き込み・削除・外部送信権限を削ぎ落とす。

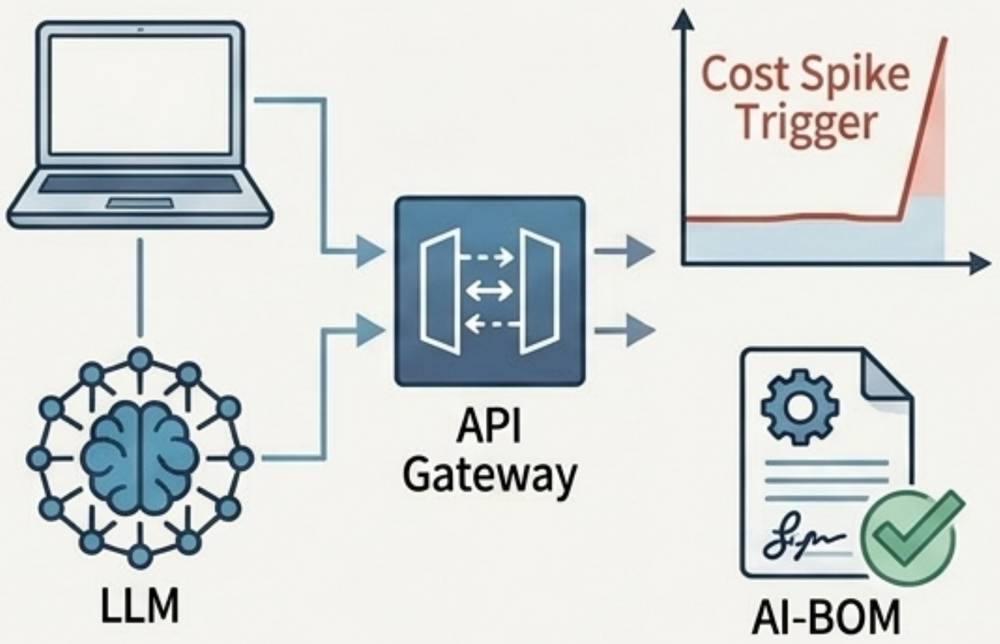
多層防衛アーキテクチャ：実務向け4軸統制モデル



単一の境界防御ではなく、OWASP、NIST等の要求事項を実務実装に落とし込んだ
「ルール」「技術」「運用」の三層構造による包括的ガードレール。

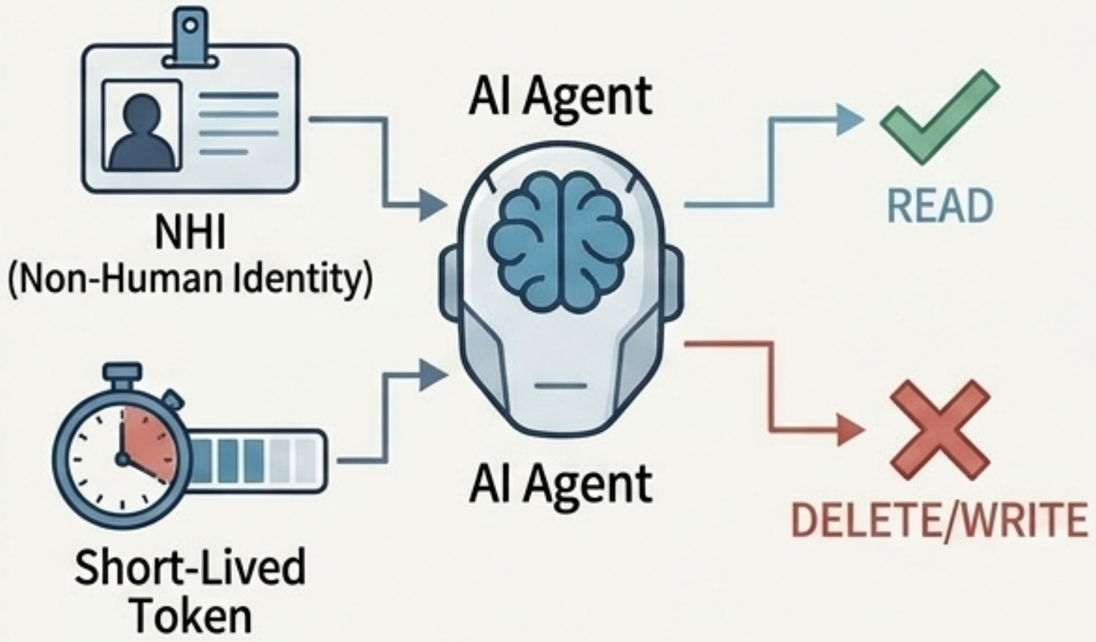
技術実装Ⅰ：可視化 (Inventory) とスコープ限定 (Scope)

第1の軸: Inventory (棚卸し)



トラフィック監視による未登録APIの検知、クラウド上のAIアカウントのコストスパイク検知。AI BOMによる動的コンポーネントの署名検証。

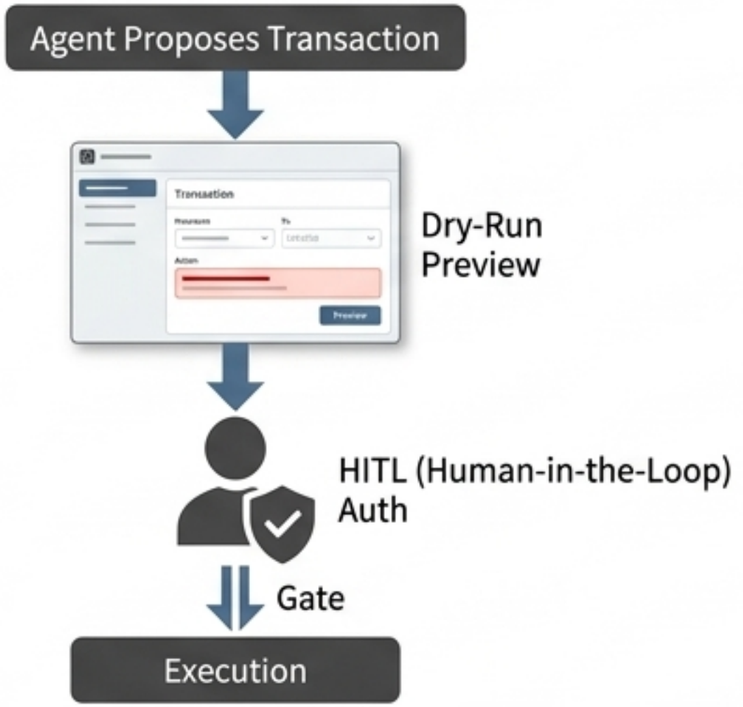
第2の軸: Scope (スコープ限定)



エージェントに固有の非人間アイデンティティ（NHI）を付与し企業のIdPと統合。セッション全体ではなく、ツール呼び出し時のみ有効な短命アクセストークンを発行。

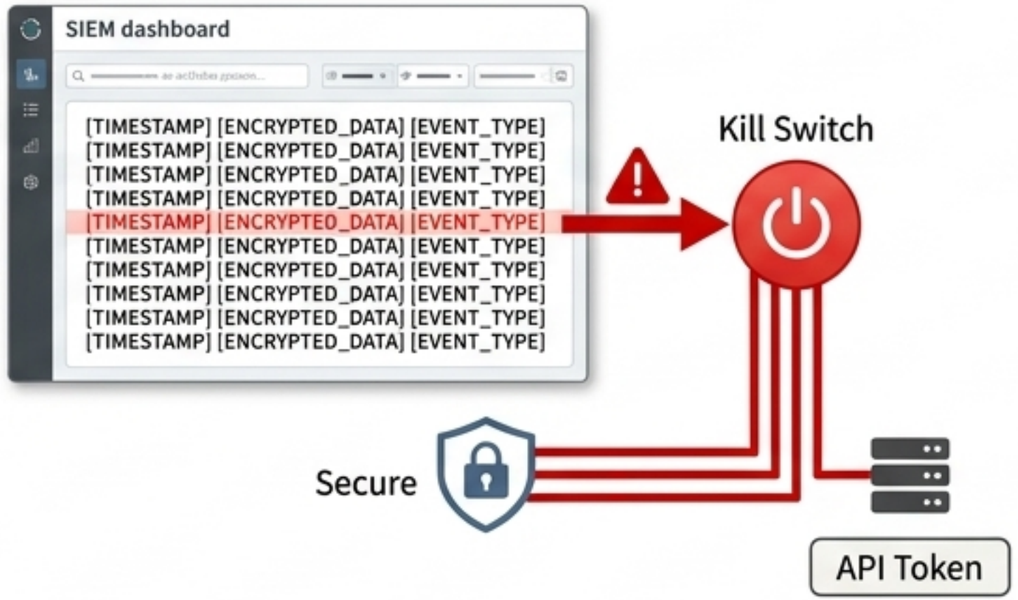
技術実装 II：人間による承認（Gate）と 監査ログ（Log）

第3の軸: Gate（人間による承認）



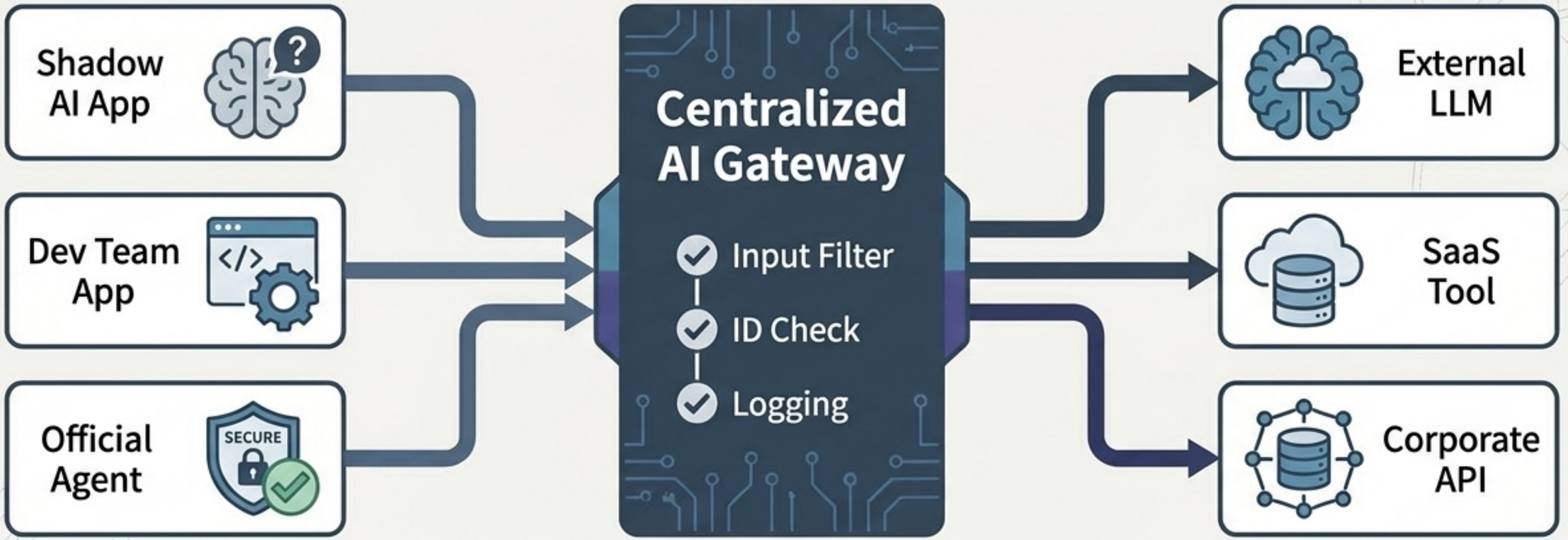
不可逆的な操作（決済、データ削除）の前に「Human-in-the-Loop（HITL)」を強制。自動化バイアスを防ぐため、実行前影響を視覚化する「ドライラン・プレビュー」とステップアップ認証を実装。

第4の軸: Log（監査ログ）



プロンプト、ツール呼び出し、人間による承認結果をマスキングしSIEMへ暗号化保存。正常ベースラインからの逸脱検知時に即座に実行を強制停止する「緊急停止メカニズム（キルスイッチ）」。

運用基盤：一元化された「AI Gateway」による自動強制



個別開発のエージェントごとにセキュリティを実装すると設定漏れ（脆弱性）が生じる。

インフラレイヤーに**AI Gateway**を配置し、すべてのアクセスを集約。プロンプトインジェクション検知、共通認証、ログ取得を「**デフォルトで** (By Default)」一括適用し、開発者の負担を下げる。

コンプライアンス：グローバル規制および国内ガイドラインとの整合

経済産業省・総務省 AI事業者ガイドライン

「自律の境界線」の整理と人間の判断介在（Axis 3: HITL）に適合。

NIST AI RMF / CSA Agentic Profile

Autonomy Tiers（自律度合い）に応じた監視要件と動態モニタリング（Axis 4: Log）に適合。

4-Axis Architecture

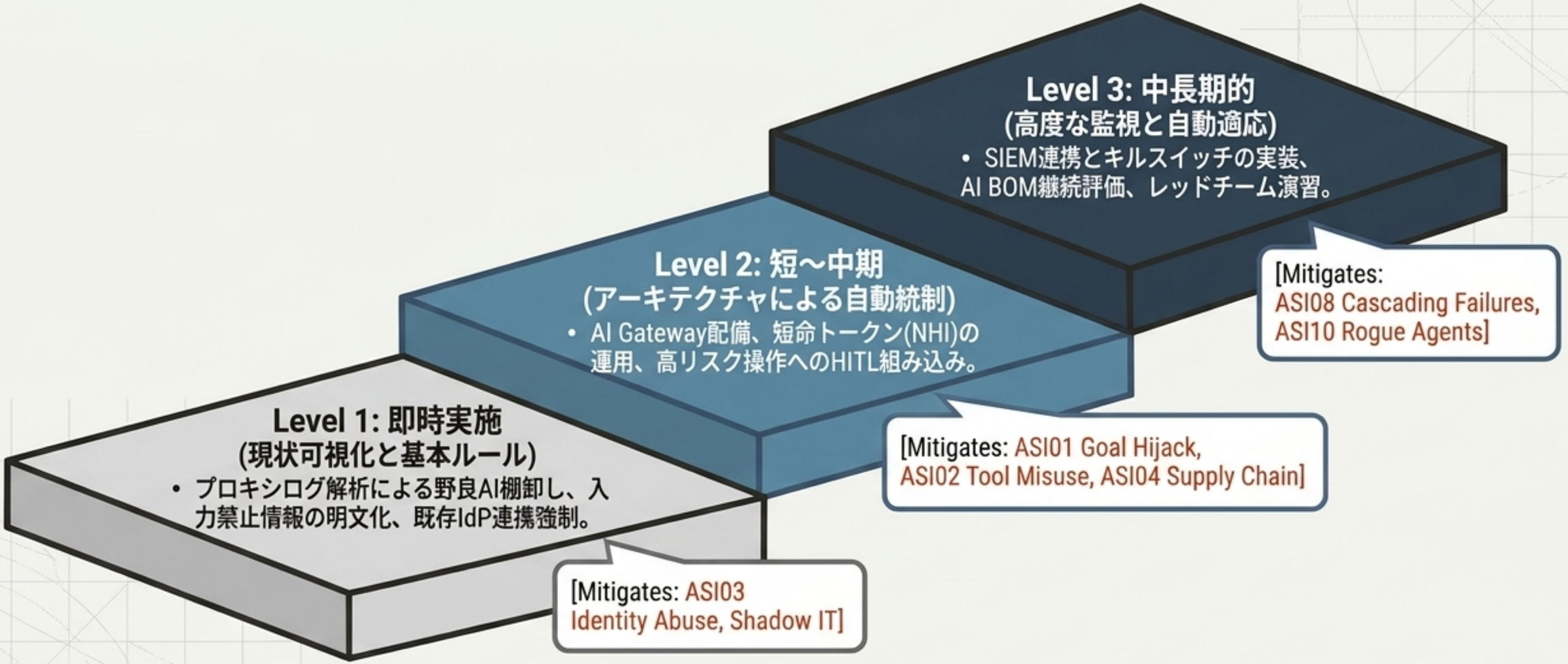
EU AI Act（欧州AI法）

高リスクAIにおける詳細なログ保持、透明性、人間による監視義務（Axis 3 ＆ 4）を網羅。

OWASP AISVS 1.0

技術層でのセキュリティ要件定義、データ完全性、アクセス制御の検証基準（Axis 1 & 2）ととして活用。

戦略的ロードマップ：安全なAI自律性のための3フェーズ



技術的ガバナンスの確立が、AIによる生産性向上の真の前提条件である。