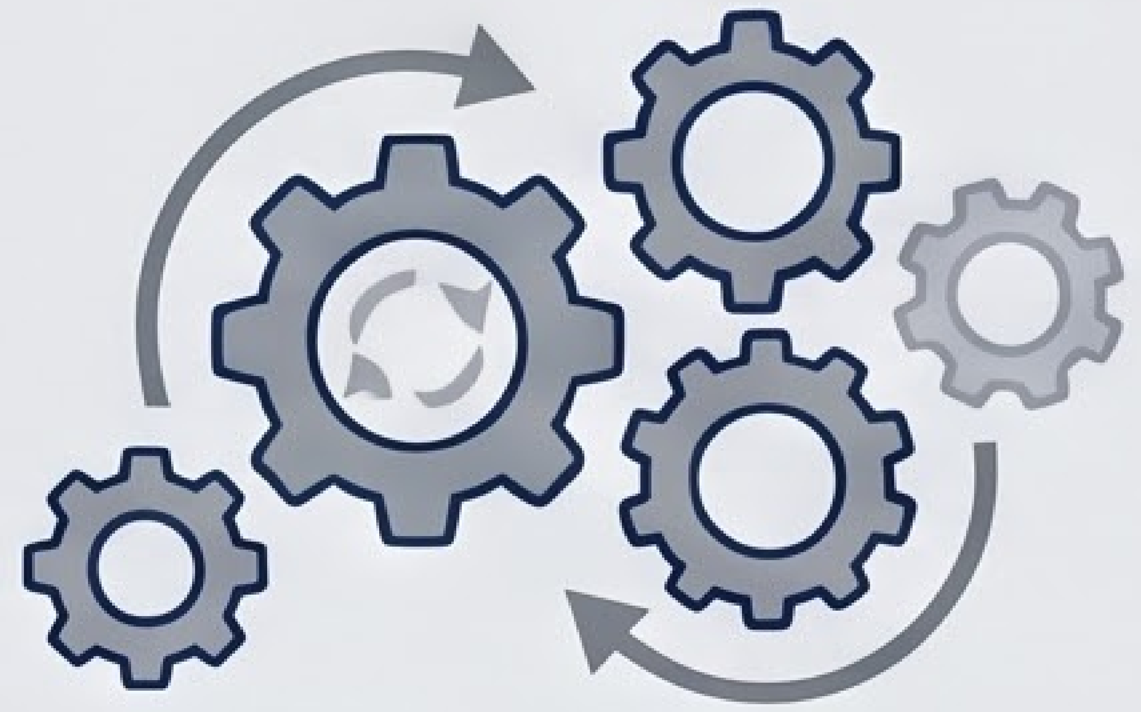
知財部門における「野良AIエージェント」対策ガイド：エージェンティック・ガバナンスの構築

自律型AIの隠れたリスクを可視化し、安全に管理・運用するための枠組み

脅威の変遷：RPAから自律型エージェントへ

RPA（受動）

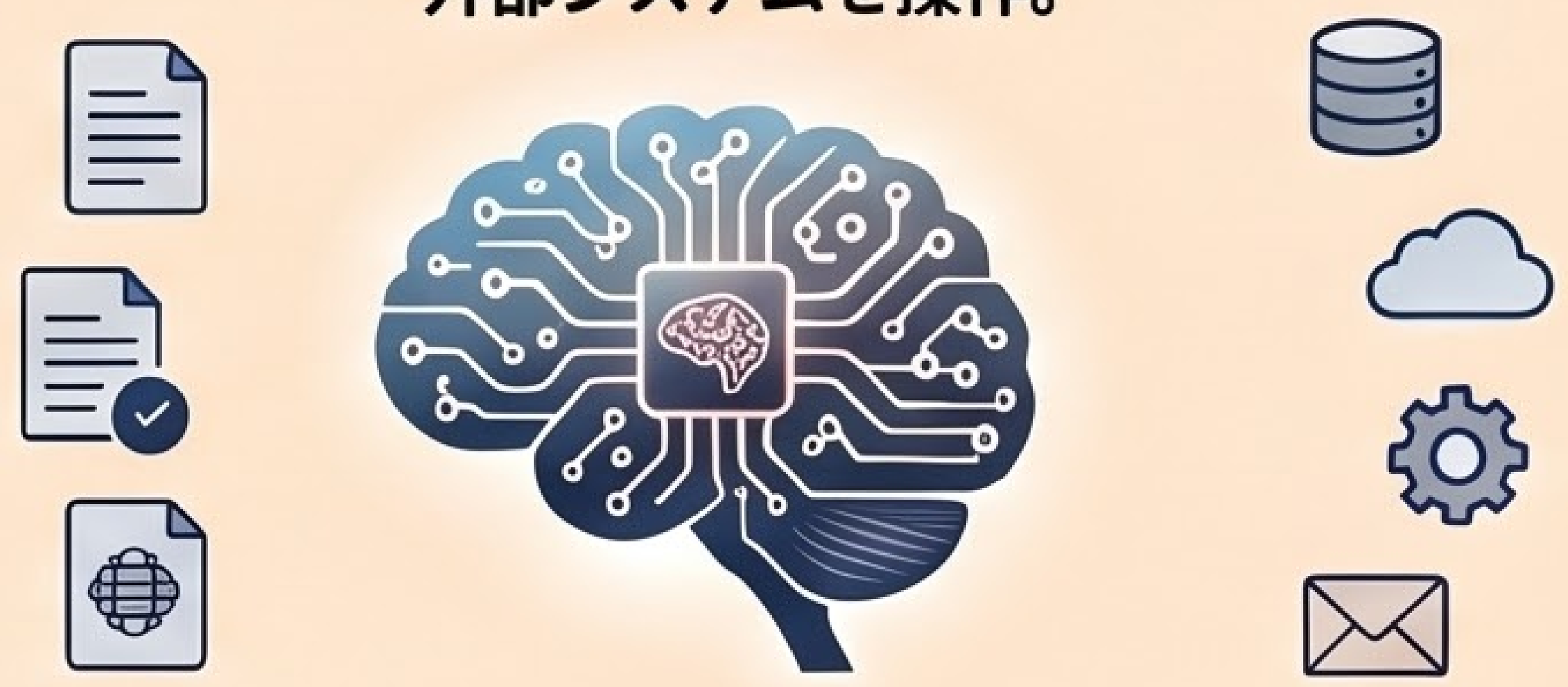
定型的な業務プロセスの自動化。



定型的な業務プロセスの自動化。

AIエージェント（能動）

高度な推論に基づき、自律的に判断して外部システムを操作。



アプリケーションの爆発的な増加
パイプコーディングにより、100個から10,000個へ激増（1,000人規模企業）

脆弱性の正体：文脈層の攻撃と影のネットワーク

Model Context Protocol (MCP)：知財部門特有の有効的投分出の認来。
シャドーネットワーク：未開特待やM&A極線約書など、最重要破が外遣出するリスクに遁載。

リスクベースの監督モデル（ガバナンスの核心）

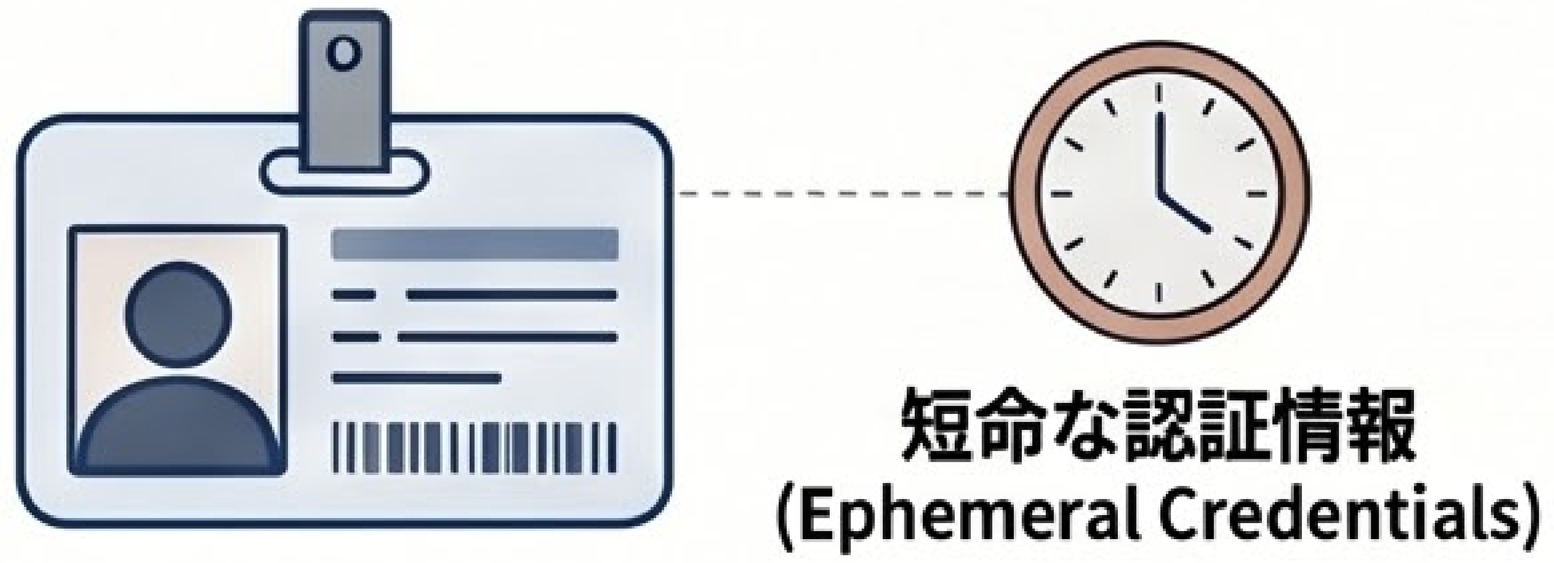
リスクに応じた「人間の介在」の設計

リスクレベルの監督会法 → 監督寠のモデルを監督介入 → 定穲案務のレート覫在

リスクレベル (Risk Level)	知財業務のユースケース例	監督モデル (Oversight Model)	必須となる技術的統制 (Technical Controls)
High Risk	契約書の作成と承認、NDA締結、特許出願の提出	Human-in-the-Loop (HITL)：明示的な人間の承認が不可欠	UI承認、厳格なRBAC、一時的なトークン(Ephemeral tokens)
Medium Risk	先行技術調査の草案、クレームマッピング、競合監視	Human-on-the-Loop (HOTL)：人間がリアルタイムで監視・介入可能	リアルタイムダッシュボード、異常検知アラート、キルスイッチ機能
Low Risk	文書の分類、知財文献の書式調整、公開特許のスクレイピング	完全自律型 (Fully Autonomous)：直接的な監視なしに動作	日次のパッチ監査、レート制限、予算上限設定

エージェンティンティック・ガバナンスの3つの柱

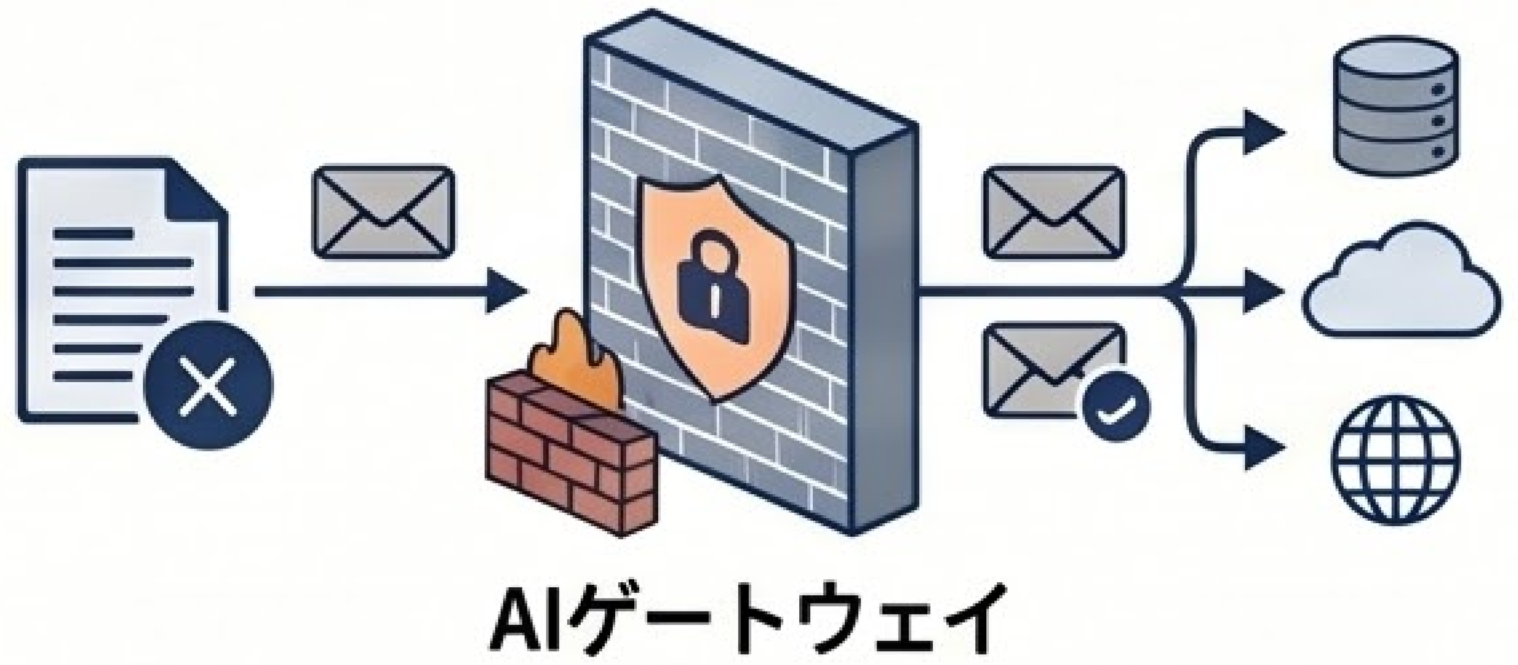
非人間アイデンティティ管理（IAM）



全エージェントに一重のデジタルIDを付与し、特定のタスク中のみ有効な動的に割り当て。

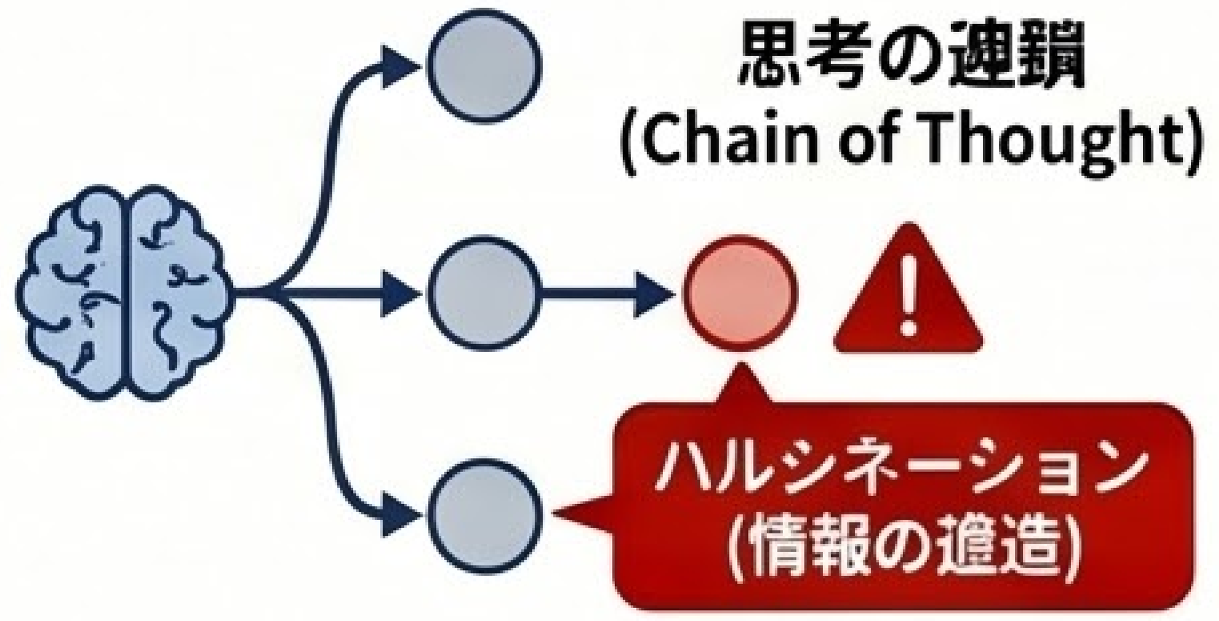
IT部門管理境界

通信経路の完全統制（AIゲートウェイ）



直接の外部接続を禁止し、社内ゲートウェイを経由させることで、個人情報や機密データの漏洩をリアルタイムで検出・マスキング。

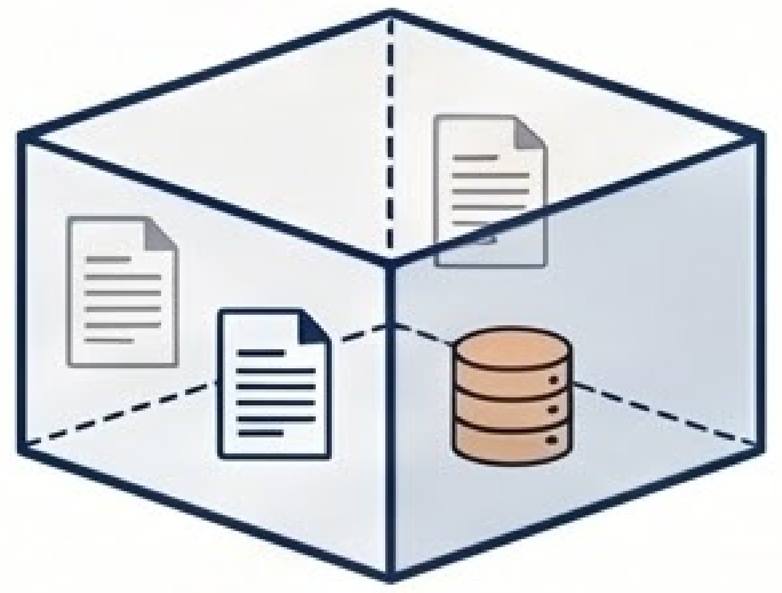
推論の可視化：AgentOps



エージェントが結論に至った思考の過程を
トレースし、ハルシネーションを早期発見。

実装へのロードマップ：フュージョンチームの結成

サンドボックスでのシミュレーション



本番環境接続前に、匿名化されたダミーデータ環境で、エッジケースに対する挙動を検証。

知財担当者とIT部門の共同開発（フュージョンチーム）

知財専門家が作成したPoCを、IT部門がセキュリティ・エラーハンドリングの観点からリファクタリングし、正式アプリへ昇格させる体制。