Strictly Confidential

Sandbox

「全面禁止」から「統制付き市民開発」へ

知財部門における自律型AIエージェントの脅威とガバナンス設計図

知財部における「野良AI」は必ず発生する。
統制の枠組みが急務である。

The Threat
（脅威の進化）

単なる対話UIから「実行主体」へ

Claude CodeやCodexは、ファイル編集・外部接続・コマンド実行機能を備えた「実行主体」。個人判断による「野良エージェント」が水面下で稼働するリスクが急増している。

The Risk
（知財特有の致命的被害）

誤答よりも恐ろしい「秘密性・権利の喪失」

高度な非定型判断を含む知財業務において、入力情報の外部送信や学習利用は「不可逆な権利喪失」に直結する。

The Solution
（実務上の正解）

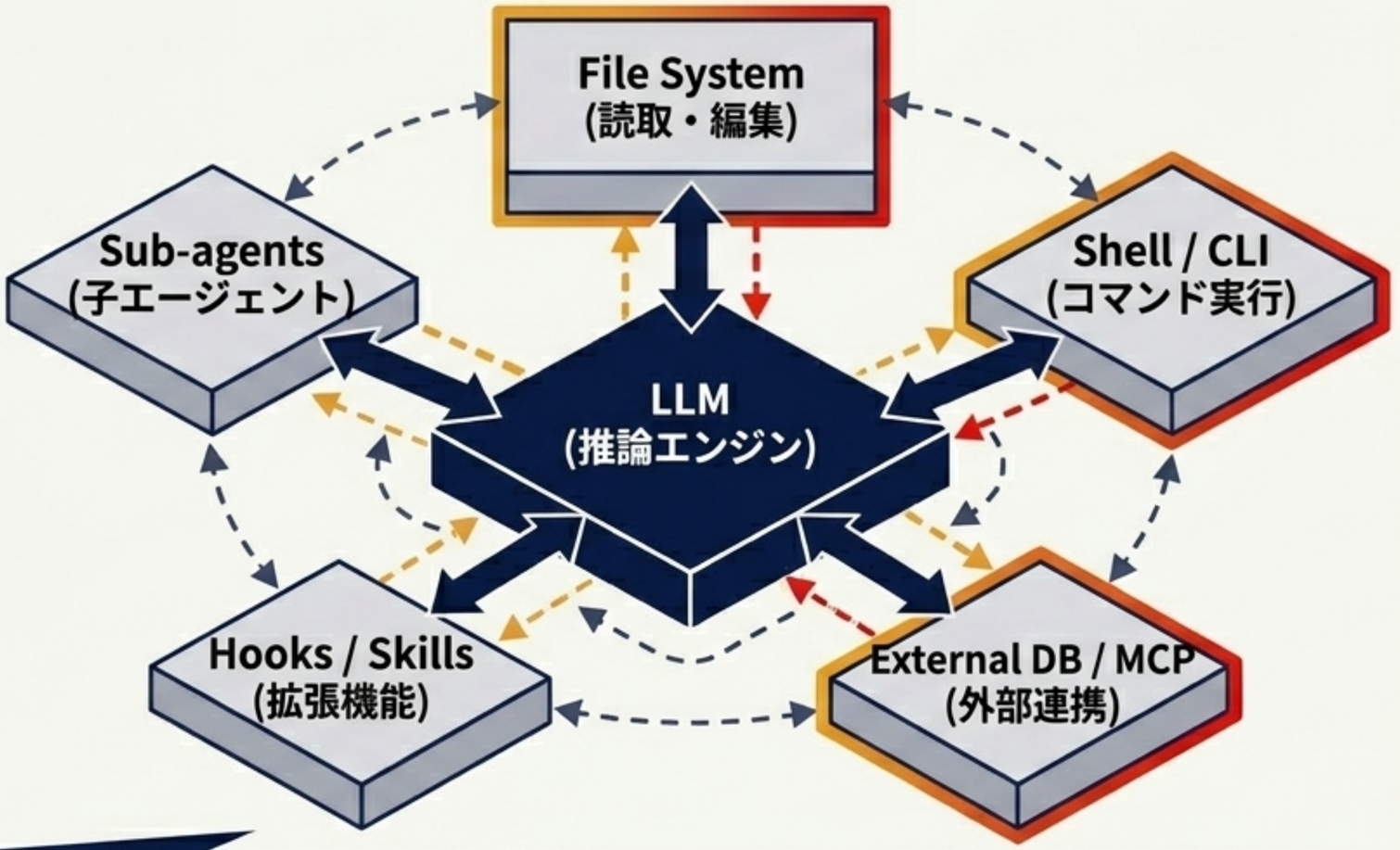
「全面禁止」ではなく「統制付き市民開発」

禁止はシャドーITを加速させる。台帳管理、承認プロセス、固有ID、Sandboxを必須化した「安全な開発インフラ」を公式に提供すべきである。

対話型AIから「自律実行エージェント」へのパラダイムシフト

Chat UI (従来の対話型AI)

Agentic AI (自律実行エージェント)

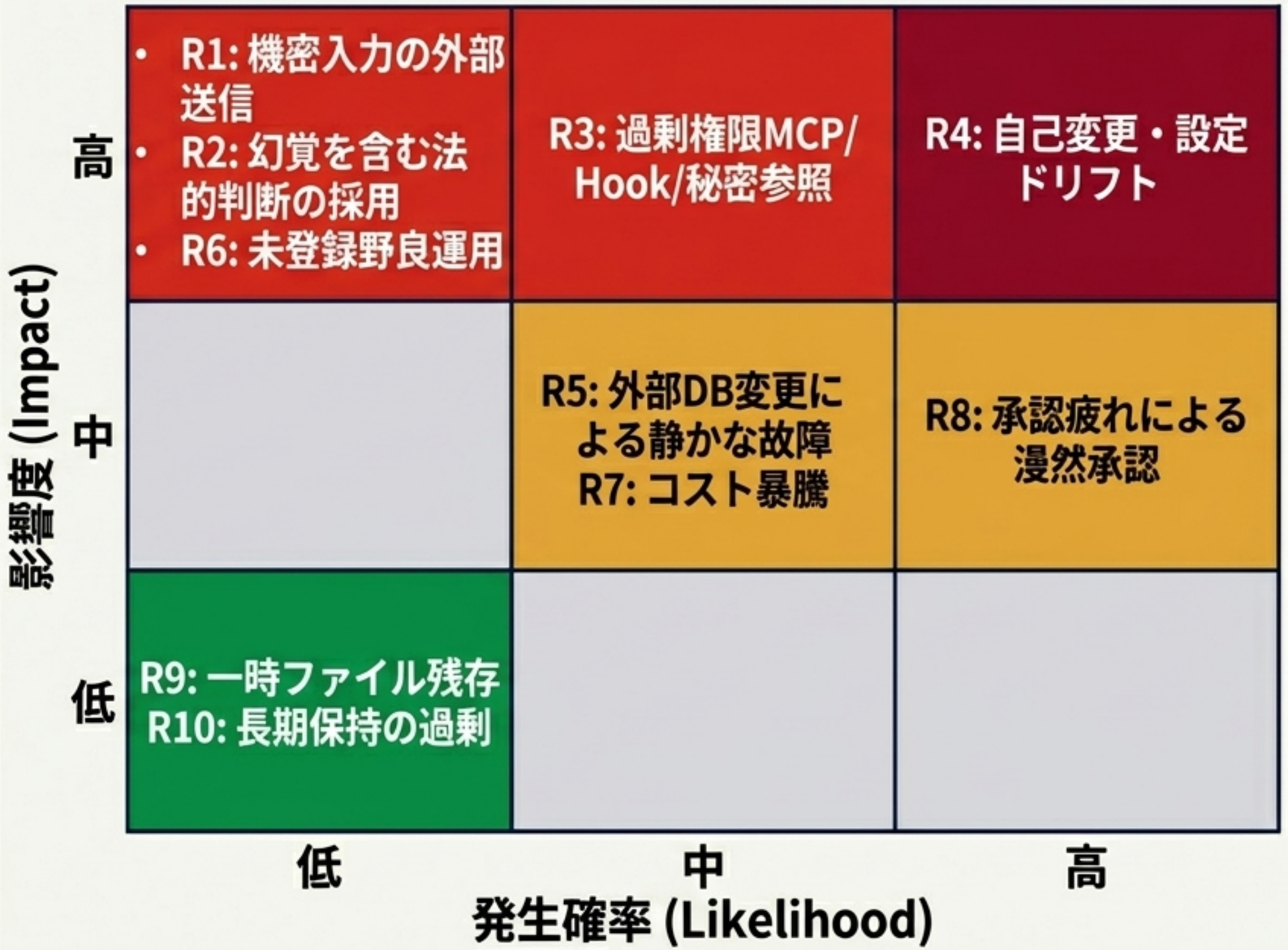


Key Takeaway: 「モデル」の統制だけでは不十分。公式仕様上、リポジトリ同梱の指示ファイルやMCPによって挙動が変化するため、「実行面全体」をガバナンスの対象とする必要がある。

野良RPAの悪夢は、より複雑な形でAIエージェントに再来する

失敗モード	RPAで起きたこと	AIエージェントでの再発形と効く統制
所在不明	未把握の夜間操作が 不正アクセス扱い	未登録CLI・個人MCPが責任不明で稼働 → 利用前登録、固有Agent ID、Owner明示
共有資格情報	人のアカウントを そのまま利用	人とAgentの区別不可 → ID分離、短命資格情報
平文秘密情報	設定にパスワード保存	.env, Prompt, Hookにトークン埋込 → Secrets Manager連携、平文禁止
静かな故障	UI変更で誤処理	外部DB仕様変更で誤要約継続 → 依存先一覧、回帰テスト
ブラック ボックス化	担当異動で意図不明	PromptやHookの意図が再現不能 → 台帳、版管理、引継ぎ
利用部門 単独開発	現場単独で作り 品質低下	高機密案件を個人設定で本番利用 → 重要案件の共同承認制

知財部における致命的リスク：誤答ではなく「不可逆な喪失」



最優先すべき脅威 (Priority Threats)

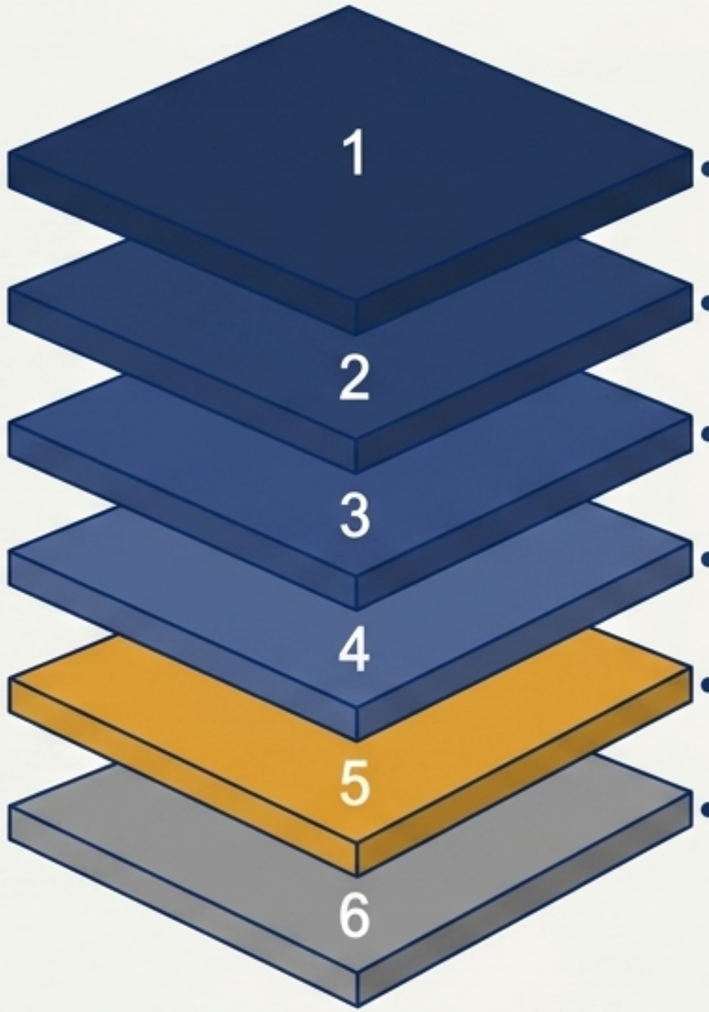
- Prompt Injection
- Insecure Output Handling
- Excessive Agency

入力前の「情報分類」が第一の統制となる。

解は「全面禁止」ではなく「統制付き市民開発（Governed Citizen Development）」

禁止だけを強めると、個人設定・個人端末・個人契約のシャドーITへ逃避する。
公式な環境を提供し、管理下に置くことが最強の防具となる。

プロンプトだけではない。統制すべき「6つの資産レイヤー」



層1: Prompt資産
対象: AGENTS.md, CLAUDE.md
ルール: コードと同格のPR承認・版管理

層2: 設定資産
対象: .claude/, .codex/, rules
ルール: write deny（自己変更禁止）

層3: 接続資産
対象: MCP, hooks.json, .mcp.json
ルール: デフォルト拒否、案件ごと申請

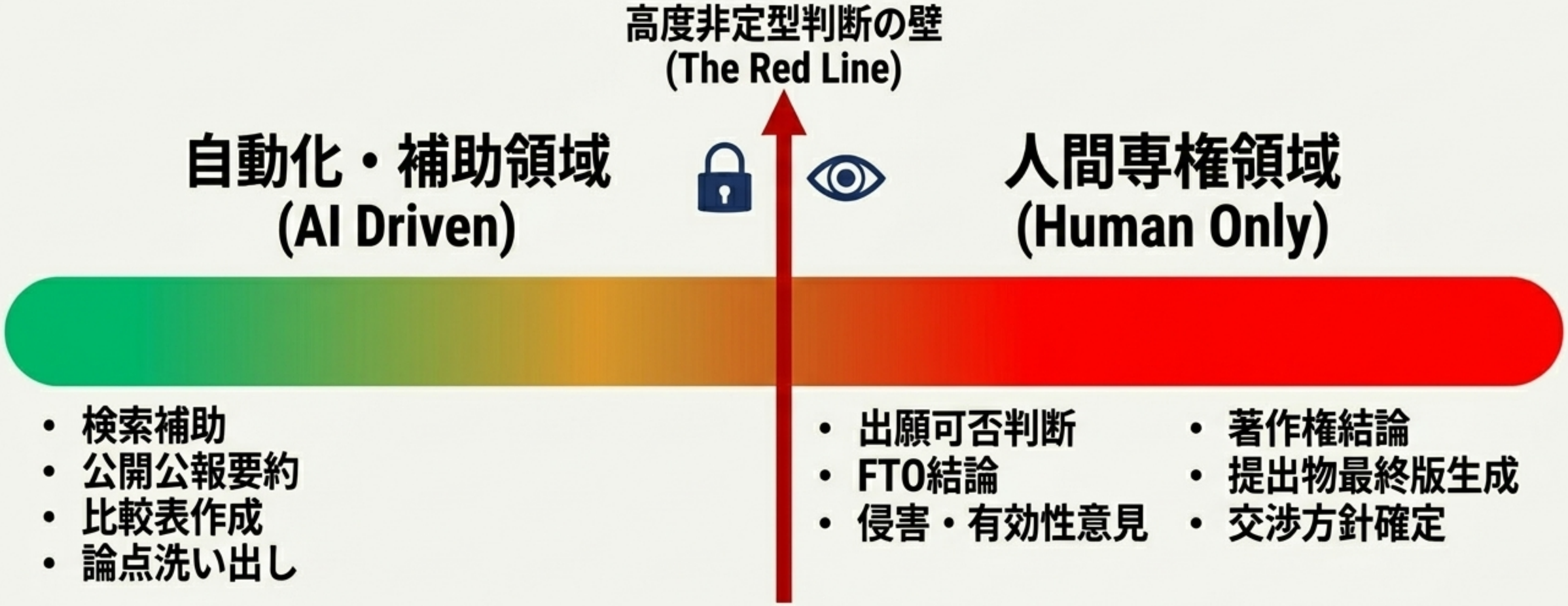
層4: 認証資産
対象: トークン, .env
ルール: ローカル保存禁止、短命資格、Secrets manager

層5: 実行境界
対象: Sandbox, 権限モード
ルール: 機密案件でのfull access / auto-approval禁止

層6: ログ資産
対象: 実行・承認ログ
ルール: セッション・エージェント相関ログの必須取得

Key Takeaway: 制御面ファイル（AGENTS.md等）はソフトウェアコードと完全に同格として扱う。

知財特有の「レッドライン（人間専権領域）」を定義する



経済産業省/特許庁/文化庁の指針に基づく。高度非定型判断は自動化対象ではなく「補助対象」。法的結論の確定と外部送信は必ず人が止める原則。

案件機密性に応じた「3層（トラフィックライト）ガバナンスモデル」

[GREEN] 許可（低リスク定型処理）

- 対象:
- 公開情報の要約、
- ダミーデータ整形
- 条件:
- 企業認証、Sandbox、
- 外部送信なし
- 運用:
- 定期再承認（180日）

[AMBER] 条件付き許可（補助作業）

- 対象:
- 未公開ドラフト構造化、
- 社内文書要約、
- 承認済み外部DB照会
- 条件:
- 知財Owner承認、
- 固有ID、Allowlist、
- ログ、二段階レビュー
- 運用:
- 定期再承認（90日）

[RED] 原則禁止・人手限定（高度判断・機密）

- 対象:
- 出願可否、
- 対外メール送信、
- 未承認MCP追加
- 条件:
- 自律実行は不可。
- 補助利用のみ。
- 最終成果物は必ず人が決裁。
- 運用:
- 個別案件ごとの法務・セキュリティレビュー

「野良化」を防ぐ中核：エージェント管理台帳（Agent Ledger）

エージェント・ライフサイクルとRACI（責任分界点）

	知財Owner	利用者	情シス	法務/セキュリティ
利用企画・案件分類	A	R	C	C
技術設計・実装	C	R	A	C
承認/例外承認	A	C	C	A
本番運用・監視	A	R	A	I
廃止	A	R	C	I

「Human-in-the-loop」を担保する承認ゲート・チェックリスト

目的と分類

- [ ] 「何をさせないか」が明確に定義されているか
- [ ] データ分類（Confidential等）が付与されているか

実行境界制限

- [ ] 法的結論の出力・外部送信がハードコードで禁止されているか
- [ ] unsandboxed状態での実行が禁止されているか

アクセスと権限

- [ ] 個人契約・個人認証基盤が排除されているか
- [ ] Egress allowlistが設定済みか
- [ ] 最小権限原則（Read-only等）が守られているか

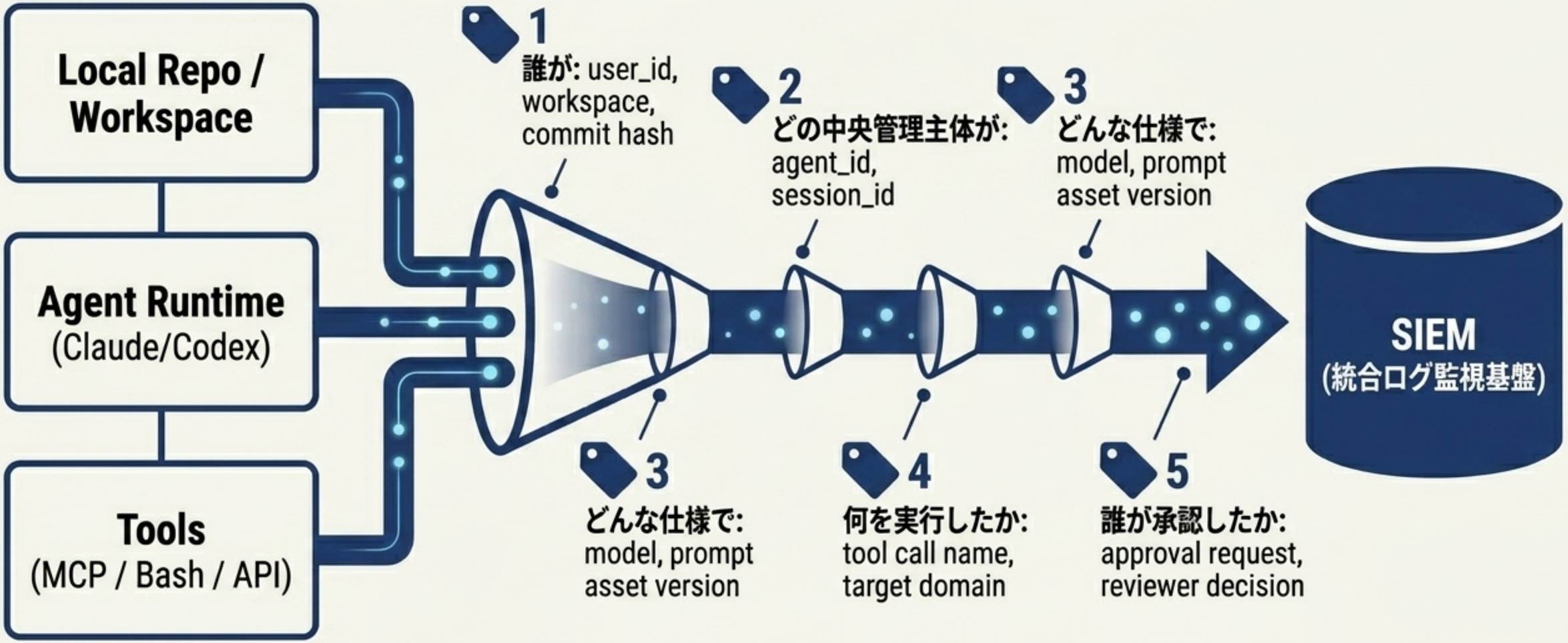
秘密情報と改ざん防止

- [ ] repo, prompt, hook内に平文の秘密情報がないか
- [ ] 制御面ファイル（.claude/ 等）への write deny が設定されているか

運用と監査

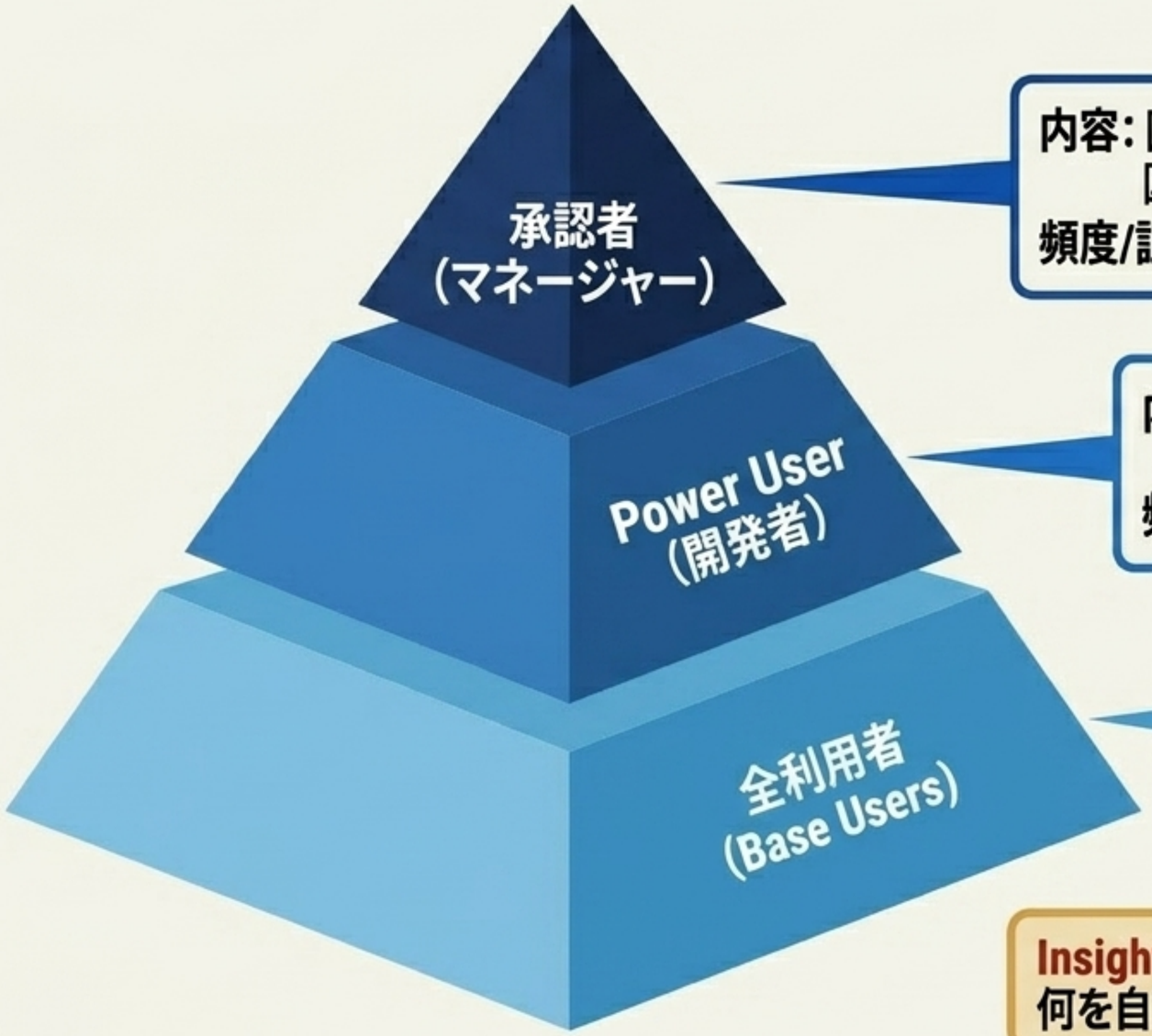
- [ ] 相関ログ（Session, Agent, User等）がSIEMに出力されるか
- [ ] 依存ツールの回帰テストが完了しているか
- [ ] 運用有効期限と棚卸日が設定されているか

サイレント・インシデントを防ぐ相関ログ・アーキテクチャ



Important: 機密Prompt全文の長期保存は避け、案件区分に応じて要約・ハッシュ・マスキングを使い分ける。

階層別リスキリング・パイプライン：技術より「境界」を教える



内容： 区分判定、例外承認、ログ確認、外部送信の妥当性判定
区分判定、例外承認、ログ確認、外部送信の妥当性判定
頻度/証跡： 初回+年1回 / ケース演習

内容： Sandbox、許可されたMCP/Hook、Secrets、Prompt版管理
頻度/証跡： 初回+半年ごと / 演習合格によりライセンス付与

内容： 情報分類、個人情報・営業秘密、禁止行為、幻覚の扱い
頻度/証跡： 初回+年1回 / 受講記録・確認テスト

Insight: 広いAIの活用方法よりも、「何を入力してはいけないか、何を自動化してはいけないか」を先に教えることが極めて重要。

導入ロードマップと追跡すべきガバナンスKPI

短期 (Low Effort)

- 暫定ポリシー発行
- 個人契約禁止
- 台帳開始
- 危険モード禁止

中期 (Med Effort)

- 企業認証強制
- Agent固有ID
- Secrets Manager連携
- Sandbox標準化
- SIEM連携

長期 (High Effort)

- Ephemeral runtime標準化
- ポリシーコード化
- 退役自動化
- 監査証跡半自動生成

ガバナンスは一度の導入で終わらない。NIST AI RMFや経済産業省ガイドラインに準拠した継続的改善のサイクルを回す。