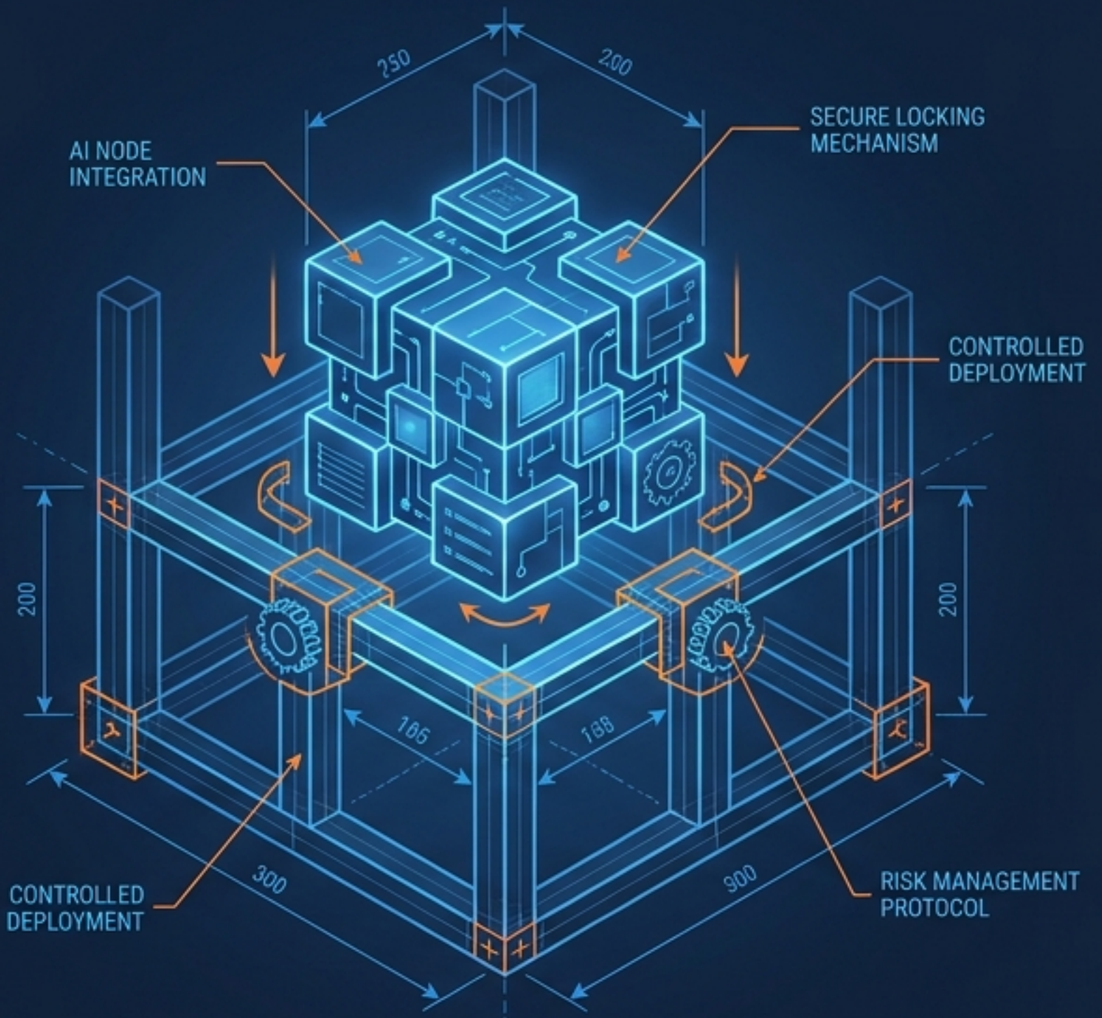
[Confidential / Internal Use Only]



知財部門で「野良AIエージェント」を発生させないための対応策

業務効率と知財リスクを両立する「IP AI Agent Management System」の構築

AIは「便利なチャットツール」ではなく、権限を委任された「業務システム」である

認識の転換

AIは自律的にデータにアクセスし、判断を補助し、外部システムへ作用するアプリケーションである。

最大のリスク

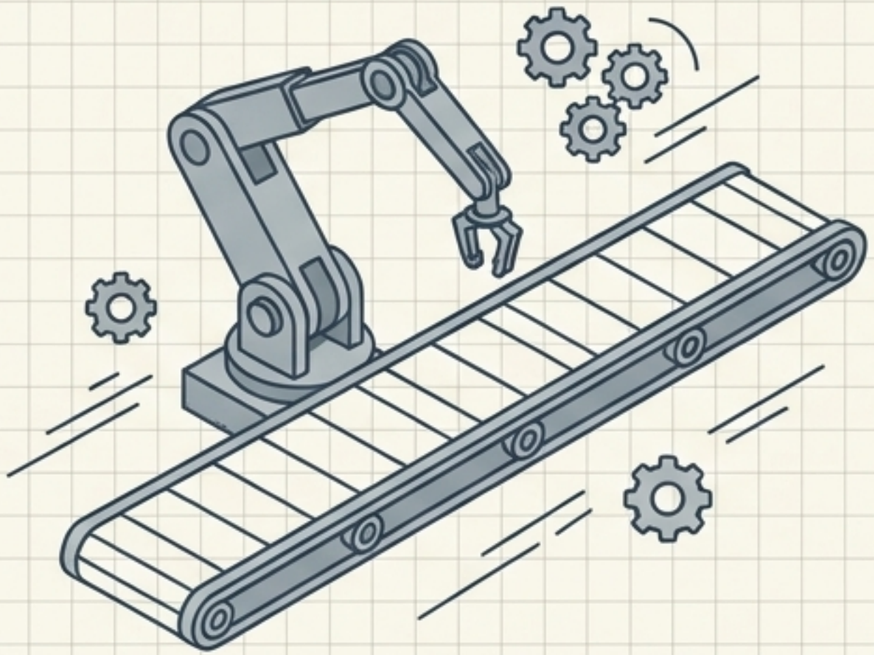
登録・監査の対象外で動く「野良AIエージェント」の増殖。

基本方針

一律の「禁止」ではなく、統制水準を定義した「管理された利用（Managed Use）」を原則とする。

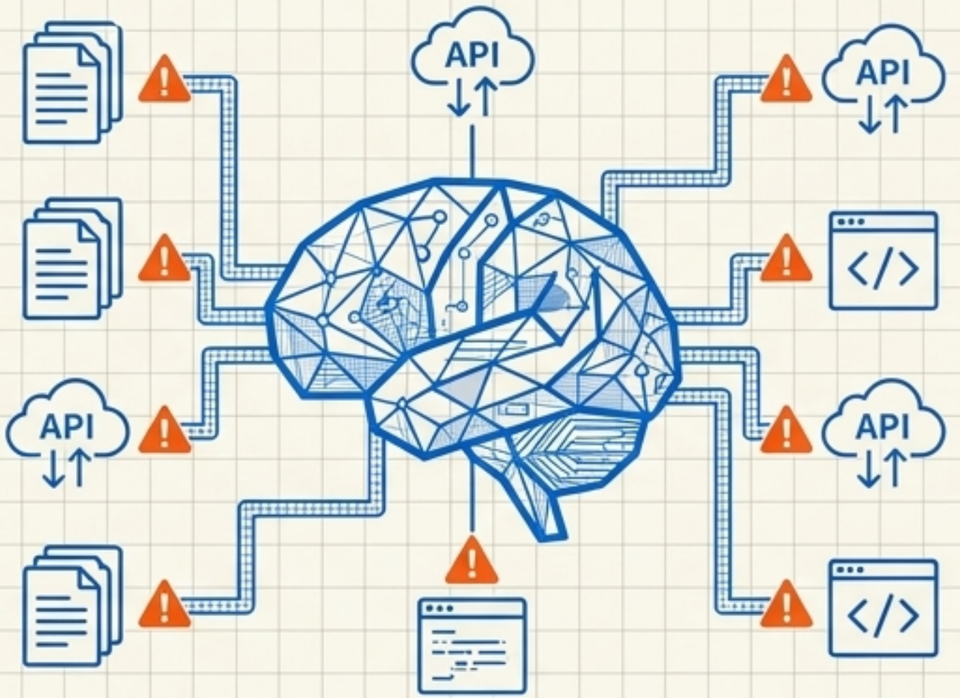
「野良ロボ」から「野良AI」へ：脅威の性質が根本的に変わる

RPAの特性



処理内容	定型処理、画面操作
主な事故	誤送信、停止
監査	ログ不足、担当者不明

AIエージェントの特性



処理内容	自然言語による推論、文書・コード生成、外部API実行
主な事故	幻覚、誤った法的示唆、秘密情報流出、権限逸脱
監査	入力根拠、プロンプト、モデル版が不明で説明責任を果たせない

なぜ知財部門は「野良AI」に対して特に脆弱なのか？

新規性喪失・NDA違反

未公開情報を外部AIに入力することで発生する秘密管理性の低下。

権利範囲の欠陥・禁反言

AI生成のクレーム案やOA応答を未検証で採用することによる権利化の失敗。

輸出管理・外国移転リスク

外国所在のAIモデルに技術情報を送信することによる外為法・EAR等の規制違反。

知財業務におけるAIエージェント・リスク分類マトリクス

	リスク分類	典型シナリオ	Major Impact（主な影響）
●	秘密情報の漏えい	未公開発明の外部AI入力	新規性喪失、競争優位喪失
●	誤生成・幻覚	存在しない判例や文献の提示	誤った出願・中間対応判断
●	権限逸脱	特許管理システムへの過大アクセス	情報の過剰取得・誤更新
●	ブラックボックス化	プロンプト・RAGが不明なまま利用	担当者異動後に保守不能
●	第三者IP侵害	出力に他人の著作物やOSSが混入	権利侵害、契約違反

パラダイムシフト：AIエージェントの解剖学

1	誤った認識
	野良化対策の失敗は、AIを「個人の工夫」や「プロンプト集」とみなすことから始まります。

2	真実の姿 (NIST AI RMF / ISO/IEC 42001準拠)
	実際には、AIエージェントは以下の管理を必要とする統合された業務システムです。 • IT資産管理 • ID管理 • 内部統制 • 品質管理

業務を守り、活用を促す「7層の防御フレームワーク」

7 継続的監視と廃止 (Audit & Lifecycle)

6 品質保証・レビュー証跡 (Quality & Review)

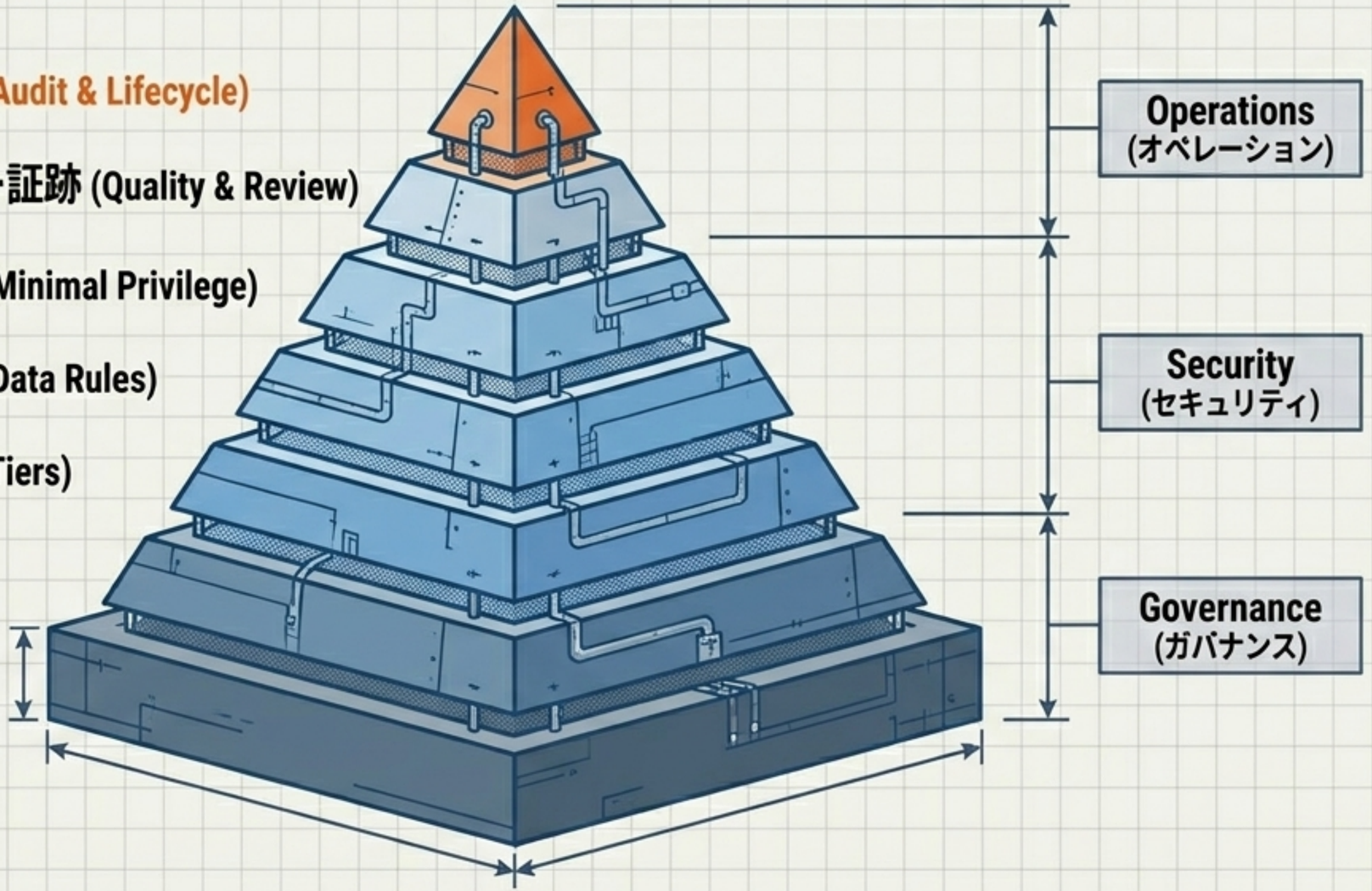
5 権限最小化と承認 (Minimal Privilege)

4 データ入力ルール (Data Rules)

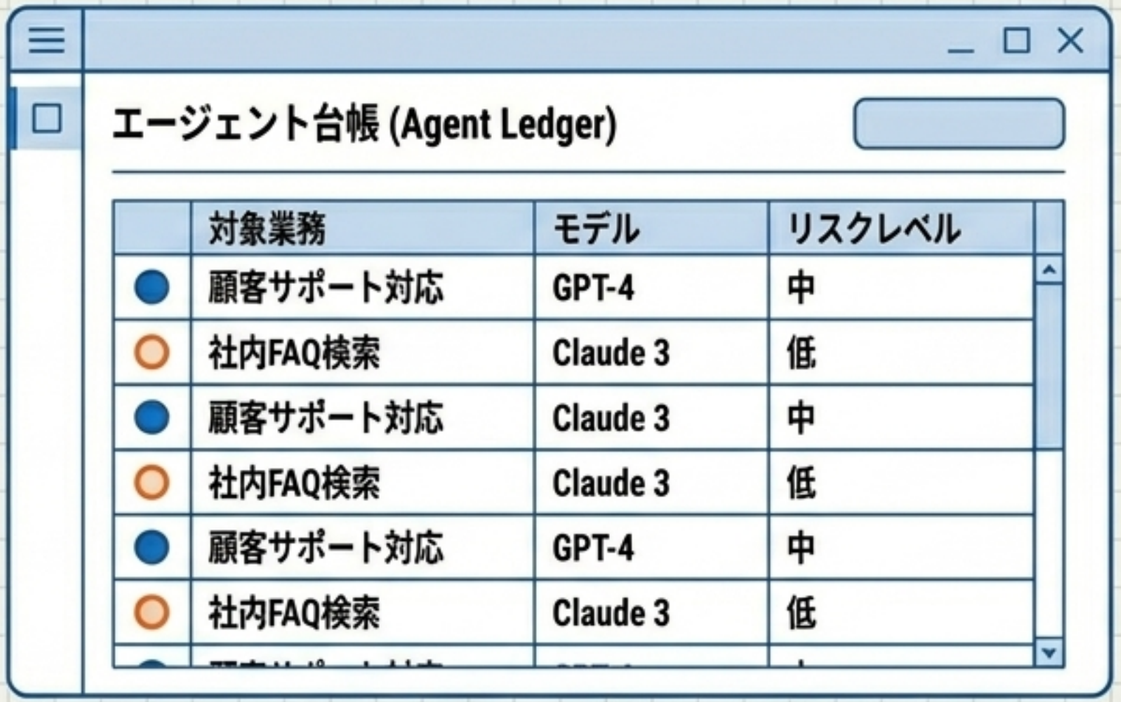
3 リスク別区分 (Risk Tiers)

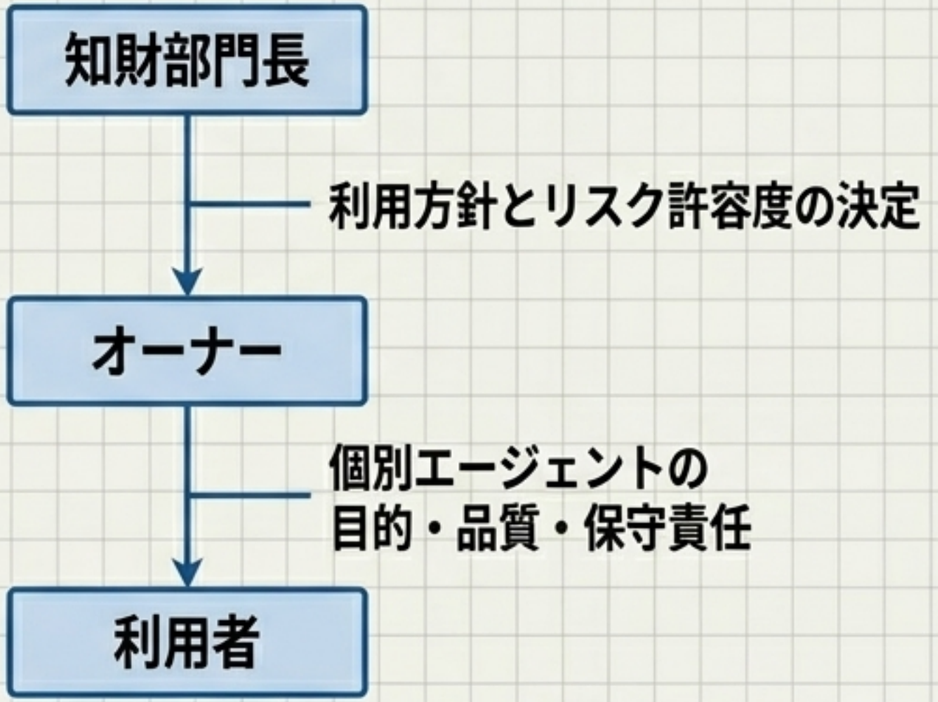
2 台帳化と可視化 (Agent Ledger)

1 方針と責任 (Policy & Roles)



第1・第2層：統制体制の確立と「エージェント台帳」



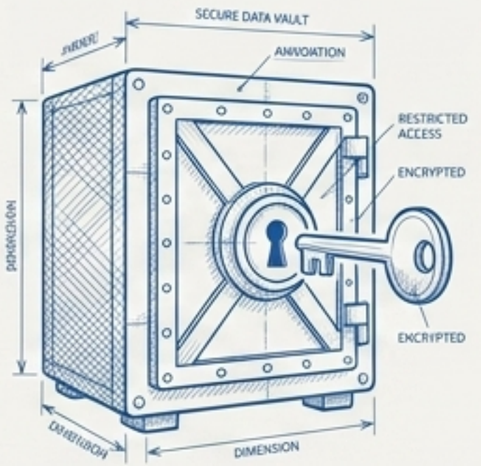


ゴールデンルール:「台帳に載っていないエージェントは、本番業務に利用できない」

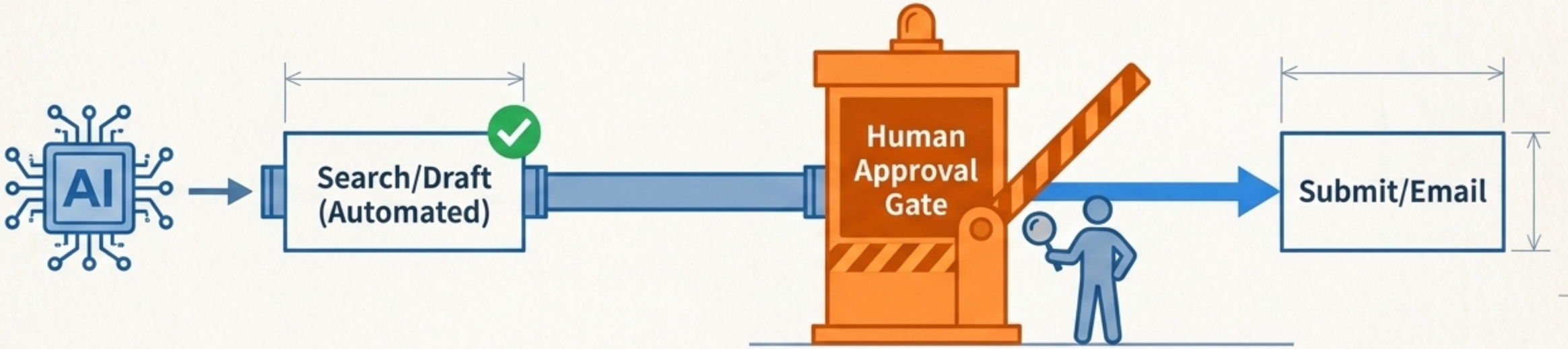
管理項目：対象業務、利用モデル、システムプロンプト、RAGデータ、リスクレベル、ログ保存場所

第3・第4層：リスクに応じた利用区分とデータ入力ルール

低リスク (公開情報)	公開特許公報、公開製品情報	登録とログのみで利用可。
中リスク (社内一般情報)	社内テンプレート、一般手順	承認済み環境と人間のレビュー必須。
高リスク (社外秘・秘密)	未公開発明、出願前明細書	個別承認、閉域環境（契約済み）、詳細ログ必須。
原則禁止 (高度秘密)	係争資料、M&A、輸出管理未審査技術	例外承認がない限り入力不可。



第5・第6層：権限の最小化と人間による「承認ゲート」



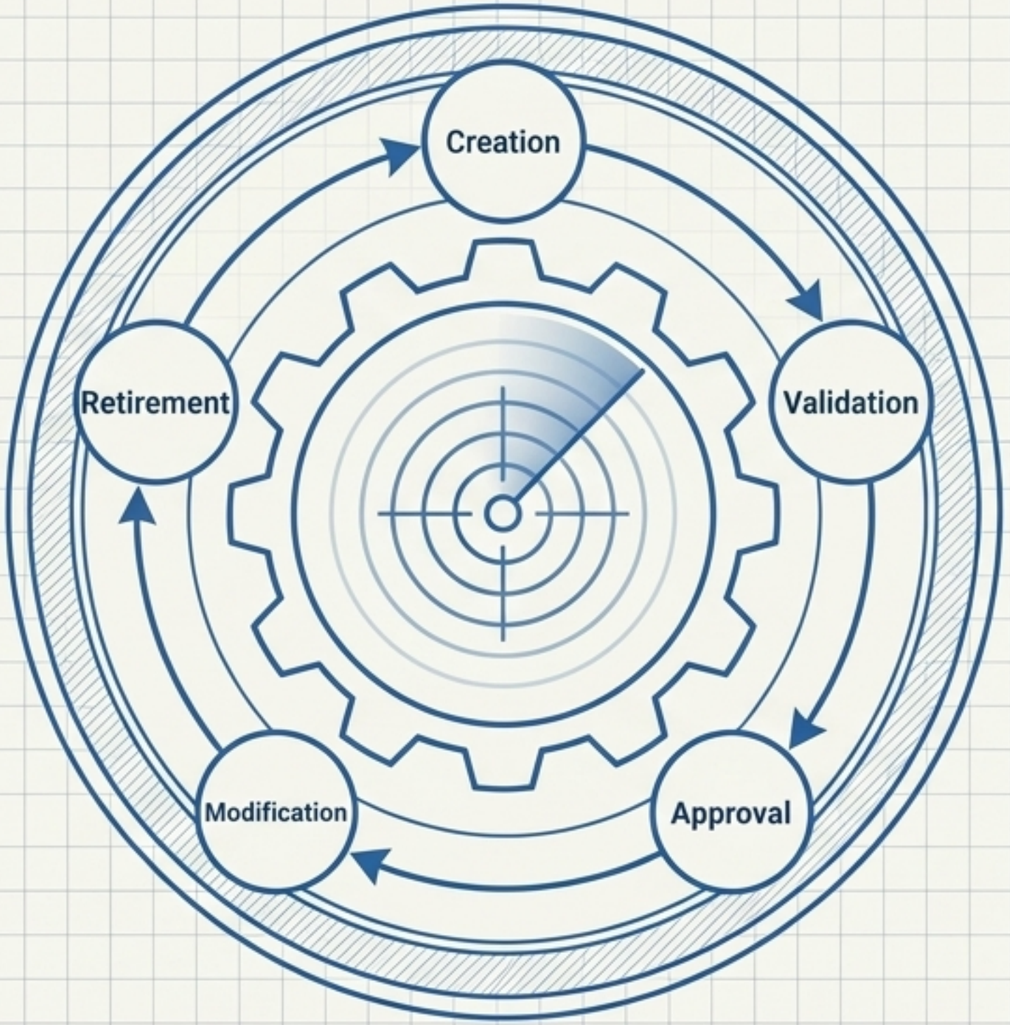
権限ルールの階層

- **検索・要約:** 許可しやすい（根拠文献の提示必須）
- **ファイル更新・ドラフト:** 差分確認・人間レビュー・版管理が必須
- **特許庁提出・外部送信:** AI単独実行は不可。有資格者・責任者の署名確認が必須。

レビュー証跡の重要性 (Review Trail Requirement)

AIの出力そのものではなく、「人間が何を確認し、何を採否したか」の証跡を残すことが専門家の責任である。

第7層：陳腐化とブラックボックス化を防ぐ継続的監査



「野良化」は運用中にも発生する。API仕様変更、担当者異動、RAGデータの陳腐化で安全なエージェントが危険化する。

監査スケジュール（Audit Schedule）	
四半期:	台帳棚卸し（未登録・所有者不明の排除）、権限レビュー
月次:	ログレビュー（禁止データ入力・外部通信の異常確認）
半期:	品質レビュー、不要エージェントの廃止確認

内製開発（Claude Code等）と外部委託におけるガードレール

内製開発ルール (Internal Dev)

- シークレット管理（APIキーをプロンプトやコードに埋め込まない）
- プロンプトの版管理と承認済みリポジトリの利用
- 本番化前のダミーデータ検証とレッドチームテスト

外部代理人・ベンダー管理 (Vendor Management)

- 委託業務でのAI利用申告の義務化
- 品質・引用確認責任は委託先に残存することの明記
- 業務終了時のデータ・ログ・ベクトルDBの削除規定

「野良化」を即座に止め、管理された活用へ移行するロードマップ

Phase 1 (0〜30日) :
把握と止血

既存エージェント・API連携の棚卸しと「暫定台帳」の作成。未公開情報の外部入力の暫定禁止。

Phase 2 (31〜90日) :
最低限の統制導入

利用区分、入力禁止ルール、ル、権限ルールの制定。高リスク業務における人間承認の必須化。

Phase 3 (3〜12ヶ月) :
継続的改善

台帳の正式運用、委託先契約の改定、NIST/ISO基準との接続による内部監査の制度化。

Immediate Visibility ➡ Continuous Optimization

社内規程の骨組み：禁止ではなく、安全なルートを設計する

1. 目的・適用範囲

社内外のAI、RAG、
委託先利用を網羅

2. 体制と定義

オーナー、利用者、
承認済みAIの定義

利用区分とデータ入力

リスクベースの制御
とDLP

4. 権限・品質管理

最小権限と専門家による
提出前確認

5. ログ・委託先・廃止

監査権と不要エージェント
の破棄

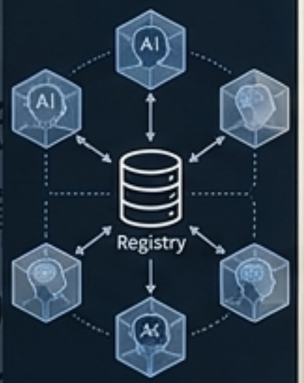
6. Policy Goal

A structural approach to provide safe routes rather than massive lists of prohibitions.

IP AI Agent Management System

1. 見える化
(Visibility)

台帳化なきAIエージェントの稼働を許さない。



2. 最小権限
(Least Privilege)

読み取り専用を基本とし、送信・更新を制限する。

3. 人間レビュー
(Human Review)

最終的な専門的判断と署名責任は人間が負う。

4. ログの保存
(Logs)

誰が、どの根拠で採否したかの証跡を残す。

5. 継続的監査
(Continuous Audit)

技術と環境の変化に合わせて定期的に見直す。

最初の一歩は、明日「現在存在するエージェントの棚卸し」を始めることです。
禁じれば潜伏します。安全な道筋を提供しましょう。