知財部における「野良AIエージェント」防止ガイドライン：統制付き市民開発の設計図

管理外AIのリスクを回避し、RPAの教訓を活かした「統制付き市民開発」フレームワーク

野良RPAの教訓：AIエージェントで再発する「失敗モード」

管理不備が招く「静かな故障」と「ブラックボックス化」

RPAで起きた「担当者不在での業務変更」や「共有アカウント利用」は、AIエージェントでも設定ファイル（AGENTS.md等）の未管理によって再発します。

「禁止」よりも「認可された市民開発」を

過度な禁止は個人端末での利用を助長するため、中央が標準ルールと環境（サンドボックス等）を整備する仕組みが有効です。

制御対象は「モデル」ではなく「実行面全体」

指示ファイル、接続設定(MCP)、Hooks、実行権限(Sandbox)をセットで真正管理する必要があります。

知財部における4大重大リスク

秘密性喪失（営業秘密・未公開発明）

一度のプロンプト入力が不可逆的な情報流出につながり、特許取得の前提となる新規性を失わせる恐れがあります。

権利取得機会の逸失と法的確実性の限界

AIによる誤った進歩性・発明該当性の判断を鵜呑みにすることで、本来得られたはずの権利を逃すリスクがあります。

第三者権利侵害

エージェントが外部DBやリポジトリから取得した情報の扱いにより、著作権侵害や営業秘密侵害を招く可能性があります。

説明責任の不能

AIが自律的に行った意思のログや承認プロセスが残っていない場合、訴訟としての法的・論理的な説明ができません。

3つの利用区分とガバナンス

Green (低リスク)

許容条件：企業認証、サンドボックス、外部送信なし

- 典型例：　公開情報の要約、ダミーデータの整形
- 承認判断：承認

Amber (中リスク)

許容条件：個別ID、Allowlist、二段階レビュー、ログ保持

- 典型例：　未公開ドラフトの構造化、承認済み外部DD開会
- 承認判断：条件付き許可

Red (高リスク)

許容条件：自律実行は禁止、人間による起業・決裁が必須

- 典型例：　出願可否の最終判断、FTO結論、対外メール法律
- 承認判断：個別設定/人手確認

人間専権事項（法的判断、対外送信）

実務ガバナンス：守るべき「8つの柱」

企業認証の徹底

機密案件での個人アカウント・個人課金利用を厳禁し、企業管理下のアカウントのみを許可します。

実行境界 (Sandbox) の必須化

実行環境を隔離し、機密案件でのフルアクセスや自動承認モードを禁止します。

人手介在 (Human-in-the-Loop)

法的結論、対外送信、権限昇格は必ず人間が最終承認するゲートを設けます。

構成ファイルの版管理

CLAUDE.mdや.mcp.jsonなどの設定ファイルをコードと同様に扱い、プルリクエスト (PR) による承認を必須にします。

接続先Allowlist

MCPや外部DBへの接続はデフォルトで拒否し、案件ごとに許可されたドメインのみ通信を認めます。

最小権限とSecrets管理

書き込み権限やネットワーク範囲を最小化し、プロンプトや設定ファイルにパスワードを平文で保存させません。

相関ログの保存

実行ログ、承認ログ、設定変更ログを紐付けて保存し、後から追跡可能にします。

定期再承認とライフサイクル管理

一定期間 (90〜100日) ごとに利用継続の必要性を確認し、不要なエージェントは廃棄します。

運用ライフサイクル・フロー：1.設計(データ分類) → 2.承認(知財・IT) → 3.実装(GR管理) → 4.検証(回復テスト) → 5.本番運用(監視) → 6.監査 → 7.変更/停止