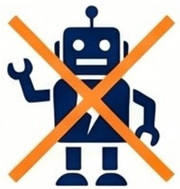
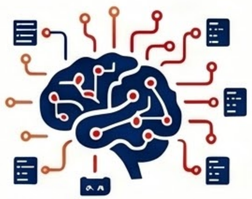
知財部門の「野良AIエージェント」対策ガイド：ガバナンスと技術統制の二層設計

野良AIエージェントがもたらす構造的リスク

野良RPAとの決定的な違い



RPA:
保守不能



AIエージェント：制御不能
- 非決定論性
- 自律的な外部接続性
- プロンプトインジェクション
- エージェント連鎖

データ流出を招く「致命的三条件（Lethal Trifecta）」

STATISTIC:
シャドーAIによる経済的損失

約19万米ドル（約2,900万円）
標準ケースと比べて事件がで
平均あたりの追加コスト

二層設計による解決策（ガバナンスと技術）

ガバナンス層：知財AIエージェントCoEの設置

- 知財・法務・ITのクロスファンクショナルな専門組織（CoE）設立
- ライフサイクル管理
- NIST AI 600-1に基づくリスクアセスメント

技術層：認証・ログ・フィルタリングの徹底

- SSO/MFAによる認証
- MCPサーバのホワイトリスト化
- DLP（データ流出防止）フィルタによる機密情報検知・ブロック

制度的根拠への準拠

- 日本のAI推進法
- 弁理士業務ガイドライン（2025年）
- EU AI Act（GPAI義務）

知財業務別・リスクと推奨制御マトリクス

	主要リスク	推奨制御
先行技術調査	ハルシネーション（捏造）、クエリによる発明内容の漏えい	HITL（人間による確認）必須、社内閉域ベクトルDB、検索ログ監査
明細書ドラフト	守秘義務違反、特許法29条1項1号（公知化）、クレーム不整合	社内RAG・閉域モデル限定、外部AP(禁止、弁理士による最終確認
FTO分析	誤った非侵害判断による経営ミス、法的助言の誤り	AIは論点抽出まで、最終判断は弁理士が行う
IPランドスケープ	競合分析情報の外部漏えい、秘密管理性の数損	送信前の社名・型番マスキング、意思決定プロセスの記録
拒絶理由応答案	善管注意義務違反、引用例の誤読	引用例本文の直接確認、ファクトチェック手順の義務化
期限管理	致命的な権利喪失（年金・PCT等）	AIはリマインダのみ、決定論的システムでバックアップ、人間が最終承認
ライセンス交渉	交渉条件の漏えい、利益相反	クライアント単位のテナント分離、利益相反スクリーニング

12ヶ月の段階的実装ロードマップ

Stage 1（即時〜3ヶ月）：棚卸しと応急処置

- 全部員アンケートとPCログで現状把握
- 機密情報の入力禁止を徹底
- 既存の危険なAPIキーを即時無効化

Stage 2（3〜6ヶ月）：CoE設立と標準化

- 専門組織（CoE）正式発足
- 情報分類別の利用可否マトリクス
- 監査ログ基盤
- 教育プログラム整備

Stage 3（6〜12ヶ月）：高度化と認証取得

- ISO/IEC 42001等の認証取得
- 高度なインシデント訓練
- ベンダー契約の法的レビュー完了（秘密保持・学習除外等）