知財部門における「野良 AI エージェント」問題への対応策

—野良 RPA の教訓を起点にした、ガバナンスと技術統制の二層設計—

Claude Opus 4.7

1. はじめに — 知財部門で何が起きているか

現場主導の業務効率化と低コード/ノーコード開発の普及により、「野良 RPA」が日本企業の重大な統制課題として指摘されて久しい。属人化、ブラックボックス化、ID/パスワード埋込みによる漏えい、開発者退職後の保守不能等の典型的失敗パターンは、業界横断的に共有されている[1]。これと類似した、しかしより深刻な構造が、いま知財部門で進行しつつある。Claude Code、Codex、Cursor、Manus、Genspark 等の AI エージェント構築ツールを用いて、個々の弁理士・知財担当者が独自のエージェントを構築し業務に組み込み始めているが、その多くは組織的な統制下にない。

知財部門が扱う情報は、未公開発明、出願戦略、ライセンス交渉、FTO 意見等、機密性が突出して高い。誤動作が特許要件喪失や守秘義務違反に直結する一方、AI エージェントは非決定論的に動作し、MCP(Model Context Protocol)経由で外界へ自律的に書き込みを行う[2,3]。野良 RPA 時代の主たる失敗モードが「動かなくなって困る」であったのに対し、AI エージェント時代の失敗モードは「意図せず動いて困る」である。本稿は、この問題に対しガバナンスと技術統制の両面から実装可能な対応策を提示する。

2. 野良 AI エージェントの構造的リスク

2.1 野良 RPA からの教訓

野良 RPA の発生要因は、(i) 現場の効率化欲求、(ii) IT 部門のキャパシティ不足、(iii) ローコード開発の民主化、(iv) 担当者の異動・退職時の引継ぎ欠落の 4 点に集約される[1]。RPA の世界では Orchestrator 型の一元管理基盤、CoE(Center of Excellence)設置、開発標準、棚卸し、ライフサイクル管理が定着している。これらの基本骨格は AI エージェントにも転用可能だが、後述する 4 つの相違点ゆえに、転用には大幅な拡張が必要となる。

2.2 AI エージェント固有の追加リスク

AI エージェントが野良 RPA と決定的に異なる点は以下である。第一に**非決定論性**。同じ入力でも結果が変わり、バックエンドのモデル差替えで挙動が突然変化する。第二に**外部接続性**。MCP や API 経由で自律的に他システムへ書き込みを行う。第三に**プロンプトインジェクション**。入力テキスト自体が攻撃ベクトルとなり、間接的プロンプトインジェクション(ファイル内コメント、API 応答、外部 Web 頁等への悪意ある指示の埋込み)を通じて、エージェントを意図せず操作可能となる[4,5]。第四に**エージェント連鎖**。オーケストレータが副エージェントにタスクを委譲する設計が普及するにつれ、責任追跡が困難になる[6]。

Simon Willison は 2025 年 6 月、生成 AI のデータ流出構造を「lethal trifecta(致命的三条件)」として整理した。すなわち、(i) プライベートデータへのアクセス、(ii) 信頼できないコンテンツへの露出、(iii) 外部通信能力の三条件が揃うと、プロンプトインジェクションによりデータ流出が成立する[7]。知財部門の典型ワークフロー(社内発明提案書→特許文献 API 取得→外部 LLM→クライアントへのメール返信)はまさにこの三条件を満たす構造であり、構造的に脆弱である。

2.3 MCP の脆弱性 — 2025 年最大の攻撃面

MCP は 2025 年の知財 DX の主戦場であり、同時に最大の攻撃面である。Zenity の調査によれば、2025 年中に GitHub 上に登録された MCP サーバは 1 万 3,000 を超え、登録速度にセキュリティチームの目録化が追いつかないと警告されている[3]。MCP 仕様自体は通信機構を定めるのみで、認証・認可・サンドボックスを強制しない設計であり、本番接続時には最小権限の原則を別途強制する必要がある[8]。実例として、Asana MCP でのテナント間データ露出、Microsoft 365 Copilot の CVE-2025-32711、Anthropic SQLite MCP サーバでの SQL インジェクション、GitHub 公式 MCP サーバのプロンプトインジェクション等が公表されている[3,8]。OWASP Top 10 for Agentic Applications 2026 は「Agent Goal Hijacking」を第 1 位に置く[9]。

2.4 シャドーAI の経済的インパクト

IBM「Cost of a Data Breach Report 2025」は、シャドーAI が関与したインシデントが標準的インシデントに比べ平均 19 万米ドル(約 2,900 万円)の追加コストを発生させていると報告する[10]。第一生命経済研究所(柏村祐, 2025 年)は「全面禁止は地下化を招くだけで、リスクレベルに応じ

たアジャイル・ガバナンスと、心理的安全性ある通報文化が機能する」と整理する [11]。知財部門でも、許可リスト方式、安全な代替環境(社内 RAG・閉域 API)の提供、情報分類別の可否マトリクスの三点セットが現実解である。

3. 拠るべき制度的フレームワーク

3.1 日本 ― AI 推進法、AI 事業者ガイドライン、弁理士業務 AI 利活用ガイドライン

2025 年は知財部門にとって「制度的根拠が整った年」である。AI 推進法(人工知能関連技術の研究開発及び活用の推進に関する法律、令和 7 年法律第 53 号)は 2025 年 5 月 28 日に参議院で全会一致で成立、6 月 4 日公布、9 月 1 日に全面施行された [12]。同法は罰則なしのソフトロー型であるが、内閣に設置された AI 戦略本部(本部長:内閣総理大臣、全閣僚構成)を司令塔とし、第 16 条で不正利用・権利侵害リスクへの対応を規定する [12,13]。内閣府は活用実態調査、安全性対策情報収集、権利利益侵害事象調査を実施し、是正は企業名公表による社会的圧力で促す設計である [13]。

総務省・経済産業省「AI 事業者ガイドライン」は 2024 年 4 月 19 日に第 1.0 版、2025 年 3 月 28 日に第 1.1 版が公表され、Living Document として継続更新される [14]。開発者・提供者・利用者の 3 類型ごとに考慮すべきリスクと対応を整理し、広島 AI プロセス国際指針の 12 項目を反映している [14]。

知財部門に最も直接的な規範は、日本弁理士会「弁理士業務 AI 利活用ガイドライン」(令和 7 年 4 月)である [15]。同ガイドラインは以下を明示している。第一に、弁理士は守秘義務(弁理士法第 30 条)を負っており、外部事業者が提供する生成 AI に秘密情報を入力する行為は、生成 AI 提供者という第三者に秘密情報を開示することになるため、守秘義務に違反するおそれがある。第二に、結果を内容の検討・精査もせずにクライアントに提供することは善管注意義務違反のおそれがある。第三に、ハルシネーションへのファクトチェックは必ず弁理士自身が行う必要がある [15]。

内閣府「知的財産推進計画 2025」(令和 7 年 6 月 3 日)は、AI 利用発明の発明者認定論点を提示し、生成 AI の開発者(学習データの選択、ファインチューニングを行った者等)や利用者(プロンプトを入力した者等)、発明の効果を確認した者を発明者に含め得るか否か、含まれる場合の類

型や判断手法、国際調和等の論点について議論を求めている[16]。これは「明細書ドラフト AI」をどう発明者・職務発明制度の中に位置づけるかという、知財部門の実務直結論点である。

3.2 国際 — EU AI Act、NIST AI RMF、ISO/IEC 42001

EU AI Act は 2024 年 8 月 1 日に発効した。GPAI(汎用 AI)モデル提供者義務は 2025 年 8 月 2 日に適用開始となり、10^{23} FLOPs 以上で訓練されたモデルに対し技術文書化、著作権ポリシー、systemic risk モデルの通知義務等が課される[17,18]。2025 年 8 月 2 日以前に上市されたモデルは 2027 年 8 月 2 日まで猶予され、Commission の強制執行権限発動は 2026 年 8 月 2 日からとなる[18]。制裁金は最大 3,500 万ユーロまたは全世界売上の 7%のいずれか高い方である。GPAI Code of Practice(2025 年 7 月公表)には Amazon、Google、Microsoft、OpenAI、Anthropic 等が署名している[19]。

NIST AI Risk Management Framework 1.0(AI 100-1 、 2023 年 1 月) は Govern/Map/Measure/Manage の 4 機能を提示する[20]。生成 AI 特有のリスクには NIST AI 600-1(Generative AI Profile、2024 年 7 月)が対応し、12 のリスク領域(ハルシネーション、情報インテグリティ、情報セキュリティ、知的財産、バリューチェーン等)に対し 200 を超える推奨アクションを示す[21]。Cloud Security Alliance(2026 年)は、AI 600-1 にエージェント委譲境界の概念が欠落していると指摘しており[6]、知財 CoE で自前の補強が必要である。

ISO/IEC 42001:2023 は世界初の認証可能な AI マネジメントシステム規格であり、PDCA ベースで Microsoft、Synthesia 等が既に認証取得済みである[22]。知財部門単独での取得は重荷でも、CoE 設計のチェックリストとして有用である。

3.3 米国の政策転換

米国は 2025 年 1 月 23 日に大統領令 14179「Removing Barriers to American Leadership in AI」によりバイデン政権下の EO14110 を取消し、安全規制から競争力路線へ転換した[23]。2025 年 12 月の州法上書き EO も含め、米国出願戦略・米国系 AI ベンダーとの越境データ移転の前提が流動的となっている。米国出願戦略・米国クライアント対応に関わる知財部門は、契約デューデリの再点検が必要である。

4. 知財業務のユースケース別リスクと推奨制御

知財業務は単一ではない。明細書ドラフト、先行技術調査、FTO 分析、IP ランドスケープ、拒絶理由応答、期限管理、ライセンス交渉支援、ポートフォリオ分析等、リスク特性と HITL(Human-in-the-Loop)要求度は業務毎に大きく異なる。以下にユースケース別の主要リスクと推奨制御を示す。

ユースケース	主要リスク	推奨制御
先行技術調査 AI	ハルシネーション、引用文献の捏造、検索クエリ自体に発明示唆を含むことによる情報漏えい	HITL 必須、ベクトル DB は社内閉域、検索ログ監査
明細書ドラフト AI	守秘義務違反、職務発明制度との整合、ハルシネーションによるクレーム不整合、特許法 29 条 1 項 1 号の公知化	社内 RAG・閉域モデル必須、外部 API 禁止、最終ドラフトの弁理士 HITL、生成ログ保管
FTO 分析 AI	誤った非侵害判断による経営判断ミス、法的助言性質	HITL 必須、AI は論点抽出と候補列挙までに限定、最終判断は弁理士
IP ランドスケープ	競合分析情報の外部漏えい、営業秘密 3 要件のうち秘密管理性の毀損	外部送信前に競合社名・型番をマスキング、意思決定の人間記録
拒絶理由応答案 AI	善管注意義務、引用例の誤読	HITL 必須、引用例本文の確認、ファクトチェック手順義務化
期限管理(年金・PCT 国内移行等)	致命的な権利喪失	AI はリマインダ生成に限定、決定論的システムでバックアップ、最終承認は人間
ライセンス交渉支援	交渉条件の外部漏えい、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

4.1 守秘義務・営業秘密との整合

弁理士法 30 条の守秘義務、および不正競争防止法上の営業秘密 3 要件(秘密管理性、有用性、非公知性)のうち、特に「秘密管理性」が外部 AI 入力で毀損する可能性が高い[15,24]。生成 AI 事業者の規約に秘密保持義務が明示されていない場合、営業秘密の入力は秘密管理性を失わせるリス

クがある。安全側に倒すなら、未公開発明情報・営業秘密はベンダー契約で秘密保持義務が明示された閉域構成にのみ入力するべきである。

4.2 特許要件との関係(特許法 29 条 1 項 1 号)

出願前発明を生成 AI に入力した場合に「公知」化するかは重要論点である。判例の傾向は秘密保持義務の合意があれば公知性は失われないとされるが、契約上の秘密保持義務が確認できない無料生成 AI への入力は、出願前公知化のリスクが残る。自社の特許戦略上、出願前発明はベンダー契約に秘密保持義務がある環境以外で入力しないことを規程化すべきである。

4.3 職務発明制度との整合

特許法 35 条の職務発明制度において、生成 AI 自体は発明者になり得ない(現状の運用)一方、生成 AI に対するプロンプト設計や効果確認を行った従業員が発明者となる場合の貢献度評価が論点となる。知財推進計画 2025 がまさにこれを論点化している[16]。社内規程で「AI 利用発明の発明者認定基準」「相当の対価」算定方法を整備する時期に来ている。

5. 対応策 — ガバナンスと技術統制の二層設計

5.1 ガバナンス側

(a) 知財 AI エージェント CoE の設置

構成は知財部門・法務・情報セキュリティ・IT・コンプライアンスのクロスファンクショナルとし、役割は、開発標準策定、申請・承認、棚卸し・廃止、インシデント対応、教育、外部ベンダー選定・契約レビューを担う。現場の弁理士・知財担当者を巻き込んだ「分散実装+中央統制」構造とする。

(b) ライフサイクル管理

開発前申請(業務目的、扱う情報分類、接続する MCP サーバ、利用モデル、HITL/HOTL 区分の申告)、リスクアセスメント(NIST AI 600-1 の 12 リスク領域に加え、知財固有の「特許要件喪失」「守秘義務」「職務発明整合」の 3 リスクを評価)[21]、CoE または経営層による承認、四半期ごとの棚卸し、退職・異動時の引継ぎプロセス。API キーは即時 revoke、稼動ロジックはレポジトリにアーカイブする。

(c) 教育・リテラシー

弁理士・知財担当者にはハルシネーション、プロンプトインジェクション、lethal trifecta を体感する研修を実施する。管理職には「禁止すれば地下化する」原理の理解と心理的安全性ある通報文化の構築を促す[11]。EU AI Act 第 4 条は AI Literacy を義務化しており、EU 業務に関わる弁理士にとっては法的要件でもある[17]。

(d) インシデント対応体制

プロンプトインジェクション疑い、機密情報の AI 入力、ハルシネーションのクライアント到達、MCP 経由の不正書き込み、API キー漏えいの 5 類型を最低カバーする。特に特許出願期限直前のインシデントは別エスカレーション経路を設けるべきである。

5.2 技術側

(a) 認証・アクセス制御

SSO+MFA、最小権限の原則、エージェント単位のサービスアカウント、人間ユーザ ID の貸与禁止、API キーのシークレットボルト管理、定期ローテーションを基本とする。

(b) ログ・監査証跡

LLM 推論ログ(プロンプト・出力・モデルバージョン・温度パラメータ)、ツール呼び出しログ(MCP 通信、ファイルアクセス、API 呼び出し)を OpenTelemetry 標準で SIEM にエクスポートする。.env ファイル・認証ストアへのアクセスは即時アラート対象とする[5]。

(c) MCP サーバのホワイトリスト化

組織として承認した MCP サーバのみ接続可とし、GitHub 公開 MCP サーバの直接利用は原則禁止とする(脆弱性事例多数)[3,8]。社内開発 MCP サーバはコードレビュー必須とする。

(d) エージェント・レジストリとモデル・レジストリ

全エージェントとその目的、所有者、接続先、データ分類、最終更新日を中央台帳化する。RPA の Orchestrator 相当を AI エージェントでも整備する。エージェントの「定期健康診断」(テストプロンプト一覧での回帰テスト)を自動実行する。

(e) DLP・機密情報フィルタリングとプロンプトインジェクション対策

入力時に未公開発明情報・出願戦略情報・クライアント名を検知してブロックまたはマスキン

グする。CASB で未承認 AI サービスへの送信を可視化・遮断する。入力サニタイゼーション、外部から取得した MCP 応答を untrusted input として扱う原則、システムプロンプトと user input の階層分離、出力フィルタを徹底する[4,5]。

6. 12 ヶ月の段階的実装プラン

Stage 1（即時〜3 ヶ月）— 棚卸しと「出血を止める」

(1) 全部員アンケート+PC ログ+CASB による野良 AI エージェントの棚卸し(期限 60 日)。(2)「未公開発明情報・出願戦略情報・クライアント名は、契約で秘密保持義務が明示された環境以外で生成 AI に入力しない」をワンページのトップ通達として発出。(3) 承認済みツールリストの暫定公開。(4) 個人アカウント紐づき API キー、共有スプレッドシート埋め込みキーの即時 revoke とシークレットボルト集約。

Stage 2（3〜6 ヶ月）— CoE 設立と標準化

(5) 知財 AI エージェント CoE の正式発足。(6) 情報分類別利用可否マトリクスの策定。(7) NIST AI 600-1 の 12 リスク領域+知財固有 3 リスクのアセスメントテンプレート整備。(8) エージェント・レジストリと監査ログ基盤の構築。(9) MCP サーバのホワイトリスト整備。(10) 教育プログラム(弁理士・知財担当者向け 2 時間×4 回、管理職向け 1.5 時間×2 回、プロンプトインジェクション体験を含む)。

Stage 3（6〜12 ヶ月）— 認証取得と高度化

(11) ISO/IEC 42001 のギャップ分析→ガイドライン整備→必要に応じて認証取得[22]。(12) AI 事業者ガイドライン第 1.1 版・EU AI Act(GPAI 義務)への対応マトリクス整備。(13) インシデント対応訓練(特許出願期限直前のインシデント、クライアント情報の AI 誤入力、MCP 経由の意図せざる書き込み)。(14) 退職・異動時の引継ぎプロセスに AI エージェント引継ぎ項目を組込み。(15) ベンダー契約レビュー(秘密保持義務、学習除外、リージョン制限、退出時データ削除、EU AI Act/GPAI 遵守表明、ISO/IEC 42001 または SOC2 Type II の保持の条項化)。

7. おわりに

野良 RPA は「動かなくなって困る」失敗を経営課題化した。野良 AI エージェントは「意図せず動いて困る」失敗を経営課題化する。知財部門は、業界特性ゆえに、その失敗が守秘義務違反・特許要件喪失・営業秘密毀損という形で外部に可視化されやすい [15,24]。日本弁理士会ガイドライン(2025 年 4 月)、AI 推進法(2025 年 9 月施行)、AI 事業者ガイドライン第 1.1 版(2025 年 3 月)、EU AI Act GPAI 義務(2025 年 8 月適用)、NIST AI 600-1、ISO/IEC 42001 ― これら 2025 年中に出揃った制度的根拠を束ね、ガバナンスと技術統制の二層構造を、12 ヶ月で最低限のベースラインとして整備することが、いま知財部門に求められている。

「禁止すれば地下化する」という構造は実証されており、現実解は禁止ではなく安全な代替環境の提供と情報分類別の可否マトリクスである [10,11]。AI エージェントの恩恵を取り込みつつ、知財実務の信頼性を守る ― この両立を可能にするのは、現場任せでも禁止令でもなく、CoE を中核とする統制と教育の継続的運用である。

参考文献

[1]

[2]

[3]

[4]

[5]

[6]

[7]

[8]

[9]

[10]

[11]

[12]

[13]

[14]

[15]

[16]

[17]

[18]

[19]

[20]

[21]

[22]

[23]

[24]

補足ノート

一般社団法人 知財・無形資産ガバナンス協会(IPIAGA、https://ipiaga.org/)は、知財・無形資産を経営資源として活用するためのガバナンス枠組みを提唱しており、2025 年 7 月 29 日に内閣府知的財産戦略推進事務局と意見交換を実施している。同協会の公開資料や日本知的財産協会(JIPA)等の関連団体ガイドラインは、寄稿時点の最新版を一次資料で確認することを推奨する。

英国 AI Security Institute、Frontier Model Forum、OECD AI 原則(2019 年採択、2024 年 5 月改訂)等の国際動向、デジタル庁「行政の進化と革新のための生成 AI の調達・利活用に係るガイドライン」(2026 年 1 月 13 日 第 2 回先進的 AI 利活用アドバイザリーボード資料)、IPA「テキスト生成 AI の導入・運用ガイドライン」、内閣府「AI 時代の知的財産権検討会」の最新動向についても、寄稿時に各機関の公式サイトで最新の正式名称・改訂状況を確認することを推奨する。

本稿は 2026 年 5 月時点で確認可能な情報に基づく。AI エージェント・ガバナンスは四半期単位で重要動向が更新される領域であり、寄稿掲載時点での直近の EU AI Act 執行事例、日本の AI 基本計画策定状況、OWASP Agentic Top 10 の更新、米国 Trump 政権 AI Action Plan の進展、特許庁・WIPO 等の AI 利用発明審査基準動向を必ず追補されたい。