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

作成者：Manus AI

作成日：2026年5月12日

エグゼクティブサマリー

知財部門におけるClaude Code、Codex、社内LLM、RAG、MCP、特許データベースAPI等を組み合わせたAIエージェントの活用は、先行技術調査、明細書ドラフト、拒絶理由通知対応、契約レビュー、ポートフォリオ分析、期限管理などの業務効率を大きく高める可能性がある。一方で、個人やチームが独自に作成したAIエージェントが、登録・承認・監査・保守の対象外で動作する場合、かつてRPAで問題となった「野良ロボット」と同様、またはそれ以上に深刻な「**野良AIエージェント**」問題が発生する。

結論として、知財部門はAIエージェントを単なる便利ツールとして扱うべきではなく、**委任された権限でデータにアクセスし、判断を補助し、文書を生成し、場合によっては外部システムへ作用する業務アプリケーション**として管理すべきである。Microsoftは、AIエージェントはデータにアクセスし、行動し、委任された権限で動作するため、組織は「何が存在するか、誰が所有するか、何にアクセスできるか、何をしているか、何を止めるべきか」を把握できなければならないと整理している。[1] この考え方は、知財部門のAIエージェント統制の中核になる。

> **基本方針**：知財部門では、「禁止」ではなく「管理された利用」を原則とする。ただし、未公開発明、営業秘密、訴訟・係争、ライセンス交渉、輸出管理対象技術、個人情報、第三者秘密情報を扱うAIエージェントについては、一般的な生成AI利用より高い統制水準を設定する。

1. なぜ知財部門では野良AIエージェントが特に危険なのか

RPAにおける「野良ロボット」は、管理部門が存在を把握していないまま動作し、システム変更やデータ変更を反映できず、業務への悪影響やセキュリティ上の穴を生むものとして説明されている。[2] 三菱総合研究所も、管理担当者が不明で保守されないRPAが放置され、メール送信やファイル操作を勝手に行う「野良ロボ」と化す問題を指摘し、一定の実装・運用基準、定期的な見直し、モニタリング、定期監査の必要性を述べている。[3]

AIエージェントの場合、この問題はさらに複雑になる。RPAは主に定型処理を実行するが、AIエージェントは自然言語で指示を理解し、調査し、要約し、判断候補を作り、コードや文書を生成し、外部ツールを呼び出す。したがって、誤処理だけでなく、**誤推論、幻覚、権限逸脱、機密情報の外部送信、根拠不明な文書生成、監査不能な自律実行**が問題になる。

観点	野良RPA	野良AIエージェント	知財部門での重大化要因
処理内容	定型処理、画面操作、ファイル操作	調査、推論、文書生成、外部API実行、コード生成	発明内容、請求項、OA対応、契約条項など判断密度が高い
主な事故	誤送信、誤入力、停止、重複処理	幻覚、誤った法的・技術的示唆、秘密情報流出、権限逸脱	一度の誤りが権利範囲、期限、係争、秘密管理に影響する
管理困難性	ロボット数、担当者、保守状況	モデル、プロンプト、RAGデータ、ツール権限、APIキー、ログ	構成要素が分散し、個人PCや開発環境に潜みやすい
監査上の問題	実行ログ不足、担当者不明	入出力、根拠、プロンプト、外部通信、モデル版が不明	特許庁提出物や代理人指示の説明責任を果たせない

知財部門は、秘密情報と公的提出物の双方を扱う。未公開発明を外部AIへ入力すれば、秘密管理性や契約上の秘密保持義務が問題になり得る。AIが生成したクレーム案や応答案を十分に検証せずに採用すれば、権利範囲の過不足、禁反言、引用文献の見落とし、期限対応事故につながる。さらに、外国所在のAIサービスやモデルに技術情報を送信する場合、外国出願許可や輸出管理の論点が生じることもある。USPTOのAI利用ガイダンスも、AIツール利用に関して、秘密保持、外国出願許可・輸出規制、電子システムのアクセス、署名責任、クライアントへの義務を明示的な論点として挙げている。 4

2. 知財部門で想定すべきリスク分類

知財部門のAIエージェントリスクは、一般的な情報漏えいやサイバーリスクにとどまらない。特許、商標、意匠、著作権、営業秘密、契約、係争、研究開発、海外出願、輸出管理が交差するため、リスクを業務文脈に合わせて分類する必要がある。

リスク分類	典型シナリオ	影響	主な統制
秘密情報・営業秘密の漏えい	未公開発明、実験データ、出願前明細書を外部AIに入力する	新規性喪失リスク、秘密管理性低下、NDA違反、競争優位喪失	入力禁止情報分類、承認済みAIのみ利用、DLP、ログ監査
誤生成・幻覚	存在しない判例、誤った引用文献、誤った法域要件を提示する	誤った出願・中間対応・契約判断	人間レビュー、根拠リンク必須、検証チェックリスト

権限逸脱	エージェントがDMS、メール、特許管理システムへ過大権限でアクセスする	誤更新、誤送信、情報の過剰取得	最小権限、読み取り専用初期設定、操作別承認、停止機能
ブラックボックス化	プロンプト、RAGデータ、モデル、ツール接続が不明なまま継続利用される	担当者異動後に保守不能、監査不能	エージェント台帳、構成管理、バージョン管理、オーナー設定
特許庁提出物リスク	AI生成文書を十分に確認せず提出する	権利化失敗、禁反言、不備通知、専門家責任	署名者責任の明確化、レビュー証跡、提出前検証
輸出管理・外国移転	技術情報を外国サーバ、外国モデル、外国人開発者へ送信する	外為法・EAR等の規制違反可能性	データ所在地確認、輸出管理審査、利用サービス制限
著作権・第三者IP	AI出力に第三者著作物やコードが混入する	権利侵害、ライセンス違反、契約違反	出力レビュー、OSS・著作権スキャン、利用条件確認
代理人・ベンダー管理	外部特許事務所や調査会社が未承認AIを使用する	委託先経由の漏えい、品質低下	契約条項、AI利用申告、監査権、再委託管理

WIPOは、生成AI利用にあたり、秘密情報をプロンプトに含めることを避けること、秘密情報を扱うAIツールへのアクセスを権限ある職員に限定すること、AIツール一覧を維持し、リスクプロファイル、秘密情報を扱うツール、禁止ツールに分類すること、スタッフポリシーとトレーニングを実装することを推奨している。[5] これは知財部門にそのまま適用できる。

3. 対応の基本設計：AIエージェントを「資産・ID・業務プロセス」として管理する

野良AIエージェント対策の失敗は、多くの場合、AIエージェントを「個人の工夫」または「プロンプト集」程度に見てしまうことから始まる。実際には、AIエージェントは、モデル、プロンプト、ツール、APIキー、参照データ、実行環境、ログ、利用者、オーナー、接続先システムから成る業務システムである。したがって、**IT資産管理、ID管理、データガバナンス、内部統制、知財品質管理を統合して管理する**必要がある。

NIST AI RMFは、AI製品・サービス・システムの設計、開発、利用、評価に信頼性の考慮事項を組み込み、AIリスクを管理するための任意フレームワークである。[6] ISO/IEC 42001は、AIマネジメントシステムを確立、実装、維持、継続的改善するための国際規格であり、AI関連のリ

スクと機会を構造的に管理し、責任あるAI利用を支えるものと説明されている。[7] 知財部門では、これらを参考に、部門独自の「IP AI Agent Management System」を設計すべきである。

管理対象	最低限記録すべき項目	知財部門での追加項目
エージェント基本情報	名称、目的、オーナー、利用部門、開発者、稼働環境	対象業務、対象法域、代理人・外部委託先の関与
モデル・プロンプト	利用モデル、システムプロンプト、主要プロンプト、バージョン	法的判断を禁止する範囲、根拠提示ルール、専門家確認条件
データ	RAGデータ、参照DB、入力可能データ、禁止データ	未公開発明、営業秘密、係争資料、輸出管理対象技術の扱い
ツール・権限	接続先API、DMS、メール、特許管理システム、権限範囲	読取・書込・送信・提出の区分、承認フロー
ログ・監査	入力、出力、外部通信、ツール実行、エラー、利用者	出願番号、案件番号、レビュー者、採否判断、提出物との対応
リスク評価	リスクレベル、承認者、評価日、再評価日	秘密情報区分、法域、期限影響、対外提出影響
ライフサイクル	作成、検証、承認、変更、停止、廃止	担当者異動時の引継ぎ、代理人変更時のアクセス見直し

4. 実務対応策：七つの統制レイヤー

4.1 第1レイヤー：利用方針と責任体制を明確にする

最初に定めるべきことは、知財部門内でAIエージェント利用を誰が承認し、誰が責任を負うかである。AI推進担当だけでなく、知財部門長、情報システム、法務、情報セキュリティ、研究開発、輸出管理、必要に応じて外部代理人管理部門を含めた小規模な統制体制を置くことが望ましい。

役割	主な責任
知財部門長	利用方針、リスク許容度、重要業務への適用範囲を決定する
AIエージェントオーナー	個別エージェントの目的、品質、保守、利用者管理に責任を負う

情報セキュリティ	アクセス制御、ログ、DLP、外部通信、脆弱性管理を確認する
法務・コンプライアンス	秘密保持、個人情報、著作権、契約条項、規制対応を確認する
輸出管理	技術情報の国外移転、外国人アクセス、クラウド所在地を確認する
利用者	入力禁止情報、出力検証、レビュー証跡保存のルールを守る

4.2 第2レイヤー：AIエージェント台帳を作る

野良化防止の第一歩は、すべてのAIエージェントを台帳化することである。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、という原則を置く。Claude CodeやCodexで作成したスクリプト、MCPサーバに接続するローカルエージェント、特許DBを検索するRAGエージェント、メール下書き作成エージェントなども対象に含める。

台帳は、単なる一覧ではなく、承認状態、リスクレベル、権限、データ分類、ログ保存場所、停止方法を含めた管理台帳でなければならない。Microsoftが推奨するように、集中管理されたコントロールプレーン、統一インベントリ、所有権、継続的な行動可視化を備えることが望ましい。[1]

4.3 第3レイヤー：リスク別に利用区分を設定する

すべてのAIエージェントに同じ重い統制をかけると、現場は迂回策を探し、かえって野良化する。したがって、リスクベースで利用区分を設定することが重要である。

区分	例	統制水準
低リスク	公開特許公報の要約、公開情報だけを用いる技術分類	登録、利用ログ、基本教育で可
中リスク	社内テンプレートを用いた明細書構成案、拒絶理由の論点整理	承認済み環境、人間レビュー、根拠提示、案件ログが必要
高リスク	未公開発明、営業秘密、係争、契約交渉、特許庁提出物に関与	個別承認、閉域・契約済み環境、詳細ログ、二重レビューが必要
原則禁止	外部汎用AIへの未公開発明入力、AIによる無承認提出・送信	例外承認がない限り不可

この区分により、現場の利便性と統制のバランスを取る。重要なのは、低リスク用途を安全に開放し、高リスク用途を明確に管理することである。

4.4 第4レイヤー：データ入力ルールを具体化する

「機密情報を入れない」という抽象的なルールだけでは、現場は判断できない。知財部門では、入力可能データ、条件付き入力データ、入力禁止データを明確に分ける必要がある。

データ区分	AI入力可否	例
公開情報	承認済みAIで入力可	公開特許公報、公開論文、公開製品情報、公開判例
社内一般情報	条件付き可	公開済み製品説明、社内標準テンプレート、一般的な業務手順
社外秘・秘密	承認済み閉域環境のみ可	未公開発明、出願前明細書、研究データ、発明者メモ
高度秘密	個別承認が必要	係争資料、侵害鑑定、ライセンス交渉、M&A、重要営業秘密
入力禁止	原則不可	第三者NDAでAI利用禁止の情報、輸出管理上未審査の技術情報、個人情報の不要入力

WIPOの資料が示すように、秘密情報をプロンプトへ含めることを避け、秘密情報を扱うAIツールへのアクセスを権限ある者に限定し、AIツールをリスクプロファイル別に分類することが重要である。[5]

4.5 第5レイヤー：権限を最小化し、危険な操作に人間承認を入れる

AIエージェントは、読み取り、検索、要約、ドラフト作成にとどまる場合と、メール送信、ファイル更新、特許管理システム更新、外部提出、チケット起票などのアクションを実行する場合でリスクが大きく異なる。したがって、権限は段階的に付与すべきである。

操作類型	初期設定	本番利用条件
検索・閲覧	許可しやすい	アクセスログとデータ分類制御が必要
要約・分類	許可しやすい	根拠文献・参照箇所の提示が必要

ドラフト作成	条件付き許可	人間レビュー、採否判断、版管理が必要
ファイル更新	制限	差分確認、承認、ロールバックが必要
メール送信・外部共有	原則人間承認	送信先、添付、内容の事前確認が必要
特許庁・外部システム提出	原則AI単独不可	有資格者・責任者の最終確認と署名責任が必要

USPTOガイダンスが示すように、AIを利用して生成された提出物であっても、提出や署名に伴う責任は消えない。[4] したがって、AIエージェントによる提出物作成支援は許容し得るとしても、提出前の人間レビューと責任者の明確化は不可欠である。

4.6 第6レイヤー：品質保証とレビュー証跡を設計する

知財業務では、AI出力を「参考情報」として扱うだけでは不十分である。出力が案件判断や提出物に反映される場合、どの出力を、誰が、どの根拠で採用したかを残す必要がある。

業務	必須レビュー項目
先行技術調査	検索式、対象DB、検索日、引用文献の実在性、関連箇所、見落とし可能性
明細書・クレーム案	発明把握、サポート要件、明確性、権利範囲、実施例との整合性
OA対応案	引用発明の認定、相違点、効果、補正根拠、禁反言リスク
契約レビュー	準拠法、定義、ライセンス範囲、改良発明、成果物帰属、秘密保持
ポートフォリオ分析	データソース、名寄せ、法的状態、スコアリング根拠、更新日

レビュー証跡は、AI利用を萎縮させるためではなく、後から説明できるようにするためである。AIの出力そのものより、**人間が何を確認し、何を採用し、何を棄却したか**を残すことが重要である。

4.7 第7レイヤー：監視、監査、廃止を制度化する

野良化は、作成時だけでなく、運用中にも発生する。担当者異動、モデル変更、API仕様変更、特許DB仕様変更、法改正、外部サービス規約変更、RAGデータの陳腐化により、当初は安全だったエージェントが危険化する。三菱総合研究所がRPAについて述べるように、実装・運用基準は技術進化に合わせて見直し、モニタリングや定期監査を実施する必要がある。[3]

監査項目	頻度	確認内容
台帳棚卸し	四半期	未登録エージェント、所有者不明、未使用エージェントの有無
権限レビュー	四半期または異動時	過大権限、退職者・異動者、外部委託先アクセス
ログレビュー	月次または高リスク案件ごと	禁止データ入力、外部送信、異常実行、失敗・再試行
品質レビュー	半期	誤生成、レビュー指摘、採用率、事故・ヒヤリハット
モデル・ツール変更確認	変更時	モデル版、API、利用規約、データ保持、データ所在地
廃止確認	半期	使われていないエージェント、古いAPIキー、不要データ

5. すぐに始めるべき実装ロードマップ

最初から完全なAIマネジメントシステムを構築しようとすると、導入が遅れ、現場が独自利用を進める可能性がある。したがって、短期、中期、長期の三段階で進めるのが現実的である。

期間	目的	実施事項
0〜30日	野良化の芽を把握する	既存AIエージェント・スクリプト・RAG・API連携を棚卸しし、暫定台帳を作る。未公開発明等の外部AI入力を暫定禁止し、承認済み利用環境を明示する。
31〜90日	最低限の統制を導入する	利用区分、入力禁止情報、レビュー必須業務、権限ルール、ログ保存ルールを定める。高リスク業務では人間承認を必須化する。

3〜6か月	管理された活用へ移行する	エージェント台帳を正式化し、オーナー、リスク評価、承認フロー、監査、教育を運用する。代理人・委託先契約にもAI利用条項を入れる。
6〜12か月	継続的改善を制度化する	NIST AI RMFやISO/IEC 42001を参考に、AIエージェント管理を部門KPI、内部監査、品質管理、情報セキュリティ管理と接続する。

6. Claude CodeやCodexで作る場合の具体的な開発ルール

Claude CodeやCodexのようなコーディングエージェントは、現場の小さな課題を素早く解決できる一方、APIキー、認証情報、社内データ、ローカルファイル、外部ライブラリ、MCPサーバなどと容易に接続できる。したがって、知財部門では、以下の開発ルールを最低限適用すべきである。

項目	ルール
リポジトリ	個人PCだけに置かず、承認済みリポジトリで管理する。所有者、README、利用目的、対象データ、禁止事項を記載する。
シークレット	APIキー、DBパスワード、特許DB認証情報をコードやプロンプトに埋め込まない。シークレット管理機構を使う。
外部通信	接続先ドメイン、API、モデル、データ所在地を明示し、未承認送信を禁止する。
依存ライブラリ	出所不明なコードを入れない。ライセンス、脆弱性、更新状況を確認する。
プロンプト	システムプロンプトと主要プロンプトを版管理する。秘密情報の取り扱い、根拠提示、人間確認条件を明記する。
ログ	入力・出力・ツール実行・外部通信・エラーを、案件番号と紐づけて保存する。ただし不要な個人情報や高度秘密の過剰保存は避ける。
テスト	ダミーデータで検証し、本番データ利用前にリスク評価を行う。重要業務ではレッドチーム的な誤

	作動テストを行う。
本番化	個人の試作から本番利用へ移す際は、台帳登録、オーナー設定、権限確認、レビュー、承認を必須とする。

7. 社外代理人・委託先への対応

知財部門のAIガバナンスは、社内だけでは完結しない。特許事務所、調査会社、翻訳会社、年金管理会社、契約レビュー支援会社、特許分析ベンダーがAIを利用する可能性があるためである。したがって、委託契約や業務指示書にAI利用条項を入れるべきである。

条項	内容
AI利用申告	委託業務でAIを利用する場合、利用目的、ツール名、データ入力範囲、保存・学習利用の有無を申告させる。
禁止入力	未公開発明、営業秘密、係争資料、個人情報等を未承認AIに入力しないことを明記する。
再委託管理	AIベンダーや外部クラウド利用を再委託・再移転として扱うかを整理する。
品質責任	AIを利用しても、成果物の正確性、法的・技術的確認、引用確認の責任は委託先に残ることを明記する。
監査・証跡	必要に応じてAI利用ログ、レビュー記録、利用環境の説明を求められるようにする。
データ削除	委託終了時の入力データ、生成物、ログ、キャッシュ、ベクトルDBの削除を規定する。

8. 推奨する社内ルールの骨子

知財部門向けのAIエージェント利用規程は、長大な禁止規程にするよりも、現場が判断できる構造にすることが重要である。以下の骨子を推奨する。

章	規程内容

目的	AIエージェントの安全かつ有効な利用、秘密情報保護、知財品質確保、説明責任の確保
適用範囲	社内外のAIエージェント、生成AI、RAG、MCP、API連携、コーディングエージェント、委託先利用
定義	AIエージェント、承認済みAI、禁止AI、秘密情報、重要知財業務、外部送信
体制	部門責任者、エージェントオーナー、利用者、情報セキュリティ、法務、輸出管理
利用区分	低・中・高リスク、禁止用途、例外承認
データ入力	入力可能・条件付き・禁止データ、匿名化、マスキング、DLP
権限管理	最小権限、読み取り専用、送信・更新・提出の人間承認、APIキー管理
品質管理	出力検証、根拠確認、レビュー証跡、提出前確認、専門家責任
ログ・監査	ログ保存、台帳、定期棚卸し、事故報告、是正処置
委託先管理	AI利用申告、禁止入力、監査権、データ削除、契約条項
教育	利用者教育、開発者教育、代理人向け周知、ヒヤリハット共有
廃止	不要エージェント停止、アクセス権削除、データ削除、台帳更新

9. 最も重要な実務メッセージ

知財部門で野良AIエージェントを防ぐために最も重要なのは、AI利用を一律に禁止することではない。禁止だけでは、現場の業務課題が残り、個人利用・隠れ利用が増える。むしろ、**安全に使える承認済みルートを用意し、危険な使い方を明確に禁止し、重要業務では人間の責任とレビューを残す**ことが現実的である。

> **知財部門のAIエージェント統制は、「見える化」「最小権限」「人間レビュー」「ログ」「継続監査」の五点に集約される。**

最初の一歩は、現在存在するAIエージェント、スクリプト、RAG、プロンプト、API連携を棚卸しすることである。次に、未公開発明や営業秘密を扱う用途を高リスクとして切り出し、承認済み環境、台帳、レビュー、ログ、停止手段を整える。これにより、AIエージェントの利便性を活かしながら、知財部門にとって致命的な漏えい、誤提出、権限逸脱、監査不能を防ぐことができる。

References

[1] Microsoft Learn, Governance and security for AI agents across the organization

[2] 日立ソリューションズ, RPA運用におけるリスクとは？ トラブルを回避するための対策も解説

[3] 三菱総合研究所, RPAを「野良ロボ」にしない

[4] USPTO, USPTO issues guidance concerning the use of AI tools by parties and practitioners

[5] WIPO, Generative AI: Navigating Intellectual Property

[6] NIST, AI Risk Management Framework

[7] ISO, ISO/IEC 42001:2023 Artificial intelligence management system