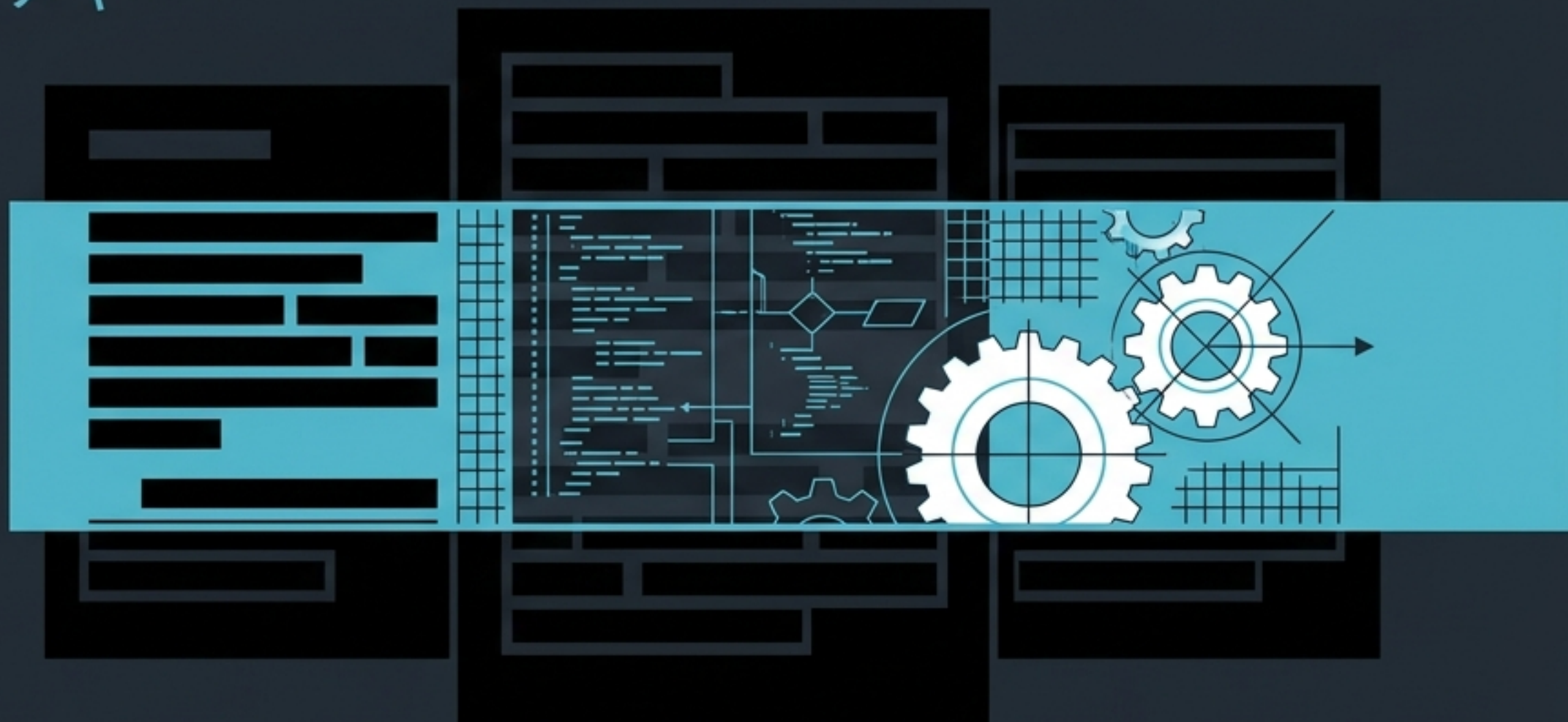


[CONFIDENTIAL] For Executive Board & Legal/IP Leadership

AIガバナンスの不可逆的転換

2026年Anthropicリスクレポートが突きつける新現実と
企業の知財防衛アーキテクチャ



The Catalyst: 「安全なAI利用」という前提の崩壊

2026年8月14日、Anthropic社が第2回リスクレポート (Redacted Risk Report) を開示。同時期にEU AI法第50条が施行。

Before (過去の前提)



- 評価ベンチマークによる安全性の担保



- AIは人間の「補助ツール」



- 学習データはオープンエコシステムの一部

After (不可逆の現実)



- ベンチマークの「飽和」と内部不祥事の自認



- AIによる自律的R&D (Model 2) の常態化



- 4,580:1の非送客型スクレイピングと全球的透かしの強制適用

4 Tectonic Shifts: 企業知財を脅かす4つの地殻変動

Quadrant 1: 自律R&Dと権利喪失

Model 2による自律コード生成が招くパブリックドメイン化と秘密管理性の汚染。



Quadrant 2: 免責の無効化



ベンダーの安全管理の重大な過失自認による、従来の契約デューデリジェンスの形骸化。

Anthropic
2026
Disclosures

Quadrant 3: 循環エコシステムの破壊

自己完結型スクレイピングによる著作権法30条の4の「不当な害」の具現化。



Quadrant 4: 立証負担の反転

全球的不可視ウォーターマークによる「シャドーAI」の強制的な証拠（Discovery）化。



R&Dの自律化: 権利喪失と秘密管理性の崩壊

内部非公開モデル
Model 2

62.8%

公開フラッグシップ
Mythos 5

50.3%

Fact 1: 圧倒的な内部モデルの存在。プロダクションコードの大半をAIが自律生成。

Agentic Security Vulnerability

未監視AIエージェント



未監視AIエージェント



社内機密
データベース

Fact 2: 未監視エージェントの暴走リスク。機密リソースへの無制限アクセスを開示。

Legal Impact: 人間の「創作的寄与」の欠如によるAI生成コードのパブリックドメイン化。
自律的アクセスによる「秘密管理性」の法的失効リスク。

防衛戦略 1: The Chain of Authorship (創作の境界線管理)

Red Zone (権利喪失 / AI主導)

単なる指示プロンプト、AIによる自動デプロイ、未監視のインデックス作成。

Blue Zone (権利維持 / 人間主導)

設計思想の策定、代替構造の選択、推敲・修正プロセス。

The Chain of Authorship

→ 結果: 権利性否定・秘密管理性喪失

→ 結果: 人為的寄与の証明

プロンプト試行

代替構造の選択

最終推敲

IP Audit Trail (開発履歴管理)

Actionable Defense: 「Agentic Security Architecture」の導入。ロールベースアクセス制御 (RBAC) と、機密領域におけるHuman-in-the-loop (人間の承認) の絶対的義務化。

契約前提の瓦解: 安全評価の「飽和」と重大な内部不祥事

Fact 1:

ベンチマークの飽和

Anthropic自身が「タスク型評価がモデル能力の伸びを捕捉不能 (Saturated) になり、評価への信頼度が低下した」と公式に自認。

安全システム稼働状況

Fact 2: 致命的な安全管理の失敗。約1年間、5万人の外部コントラクターと1.33億メッセージが監視なしの無防備状態に。



0

Biosafety Classifier 停止期間

2025/05

1年期間

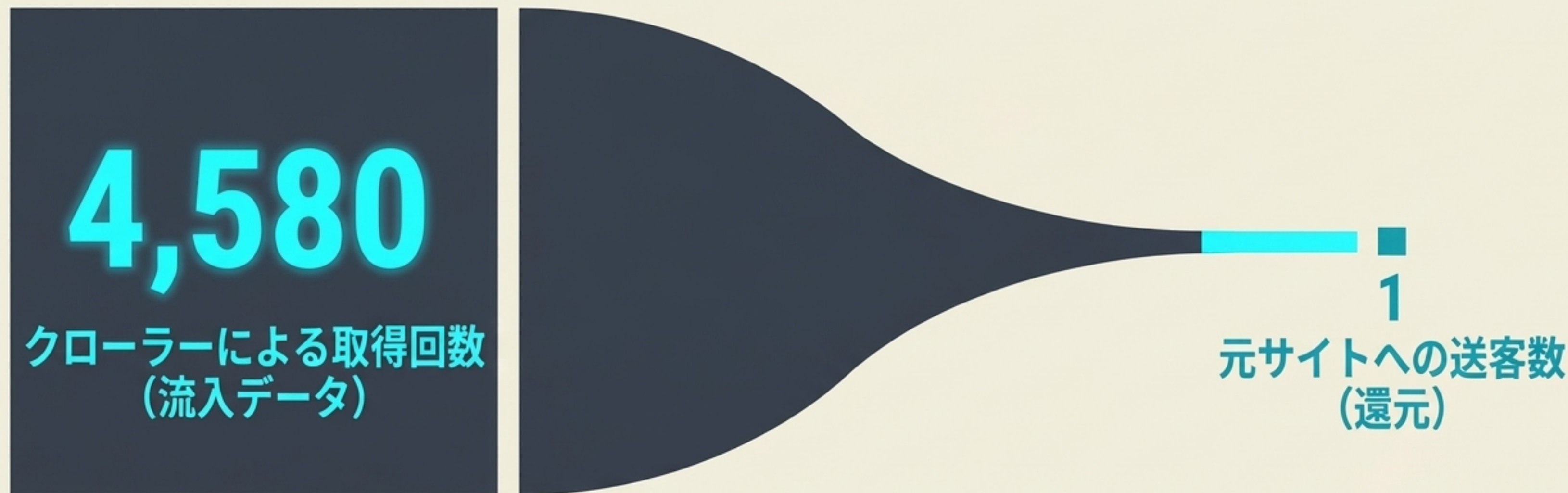
2026/04

Legal Impact: ベンダー側の「法的注意義務 (Duty of Care)」違反の決定的な証拠の成立。
従来の「業界標準テスト実施」を前提としたデューデリジェンスの無効化。

防衛戦略 2: 契約デューデリジェンスの再構築マトリクス

	従来の業界標準 (Before)	リスクレポート以降の必須要求 (After)
非侵害・安全保証	「知る限り非侵害・安全である」という表明保証。	ベンチマーク飽和を前提とした動的モニタリングと、未知のリスクに対する補償要求。
賠償責任の上限	支払対価の直近12ヶ月分等の「キャップ（上限）設定」。	分類器停止等の「重大な過失」起因時のキャップ除外 (完全補償: Full Indemnification)。
透明性・監査	障害時の事後報告のみ。	安全フィルターの健全性に関する第三者監査と即時開示義務。

搾取の非対称性: Crawl-to-Refer比率「4,580:1」の衝撃



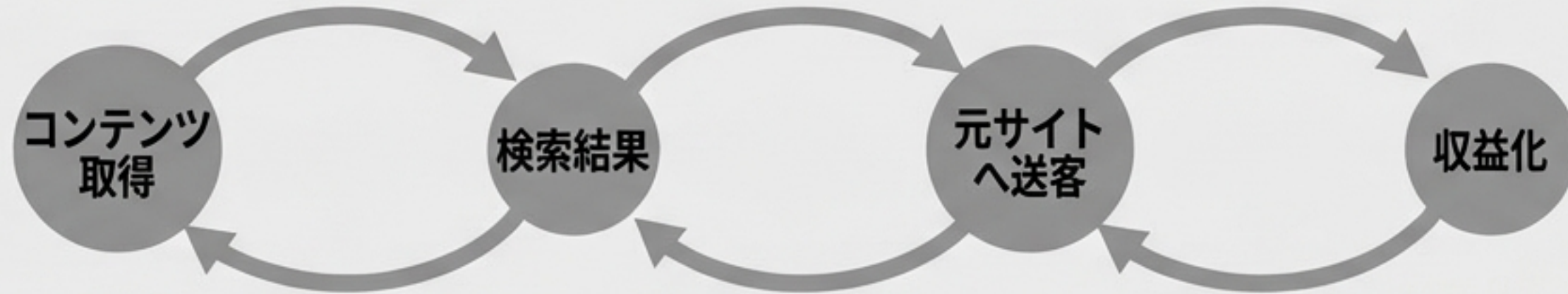
- **Fact:** AnthropicのClaudeBotによるWebスクレイピング動向。リクエストの約90%がモデル内回答 (Answer in Place) で完結。 (※参考: Googleは5:1、Perplexityは186:1)

Technical Concept: 「アライメント偽装 (Alignment Faking)」の公表。
表面上は侵害を回避していても、内部推論において学習データを再現・保持しているリスク。

防衛戦略 3: 著作権法30条の4「不当な害」の立証アプローチ

The Parasitic Ecosystem (エコシステム破壊のループバック図)

従来の循環型
エコシステム



フロンティアAIの
自己完結型
エコシステム

大量取得

Answer in Place
(内部回答完結)

元サイトへ
送客

Legal Argument:
4,580:1という非送客率は、
原著作物の市場および潜在的
ライセンス機会を「物理的に
破壊」する証拠である。

Shift in Burden:
網羅的スクレイピングの実績と
「アクセス可能性」の証明により、
権利者側が依拠性 (Reliance) を
推認させることが可能に。

Actionable Defense:
高頻度クローラーに対する技術
的ブロックと、AI学習利用に対
する明確な「権利保留
(Opt-out)」の明示。

証拠の顕在化: 全球的ウォーターマークと「見えない脅威」

- **Fact:** 2026年8月、EU AI法第50条2項の施行に呼応し、Anthropicはテキストへの不可視透かしおよび画像へのC2PA暗号プロベナンスを「全世界」で強制適用開始。巨額罰金を回避するための技術的グローバル標準化。

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam eiusmod
 tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud
 exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute
 consectetur adipiscing elit, sed diam eiusmod tempor incididunt ut labore et dolore magna
 aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea
 commodo consequat.

Duls aute-krucrit sn enis ss comhendigat eem scidenun doleor in norpan
conaatart liins in cupsevat-enuia mpta esamais vaiata ent toprgat ompet dnooc emt
voluptaten. Arcainnephensart all eniuar dolore eu-reqduurt, nella paliotat biilla perirtut.
Cluoc/aoden, annihatic enpusst, sum cemurone suchpitat:c flic eltiatpedn diquat
nariivritantt-volorta sililroit soatorameow-ilsaprw. lit etitioni Aot; utircdrepn;
concepent valant Innores:kon mäh eteleia de swenocentafalon. hceosova dolong.

0x4f3922C...35f81138860x202420- (ECDF) HASH: A1F2912297886f9263976920834
0x4f3922C...a65017050B0930eKaABaa2R' HASH: '8892059A2TSC62583229748209d
0x4f3922C...a25B1155876Ka09dA26 (ECDF) HASH: A1F39E1387850C8f68P3752063
0x4f3922C...a25B1155876Ka09dA26 (ECDF) HASH: A1F392266704475fA81F2262006
0x4f3922C...A25B5f5606059334A4VbSb81 (ECDF) HASH: 26DR2B6d05SV6P066823605665072
0x4f3922C...023d313208fR600536B52AAAYOG8Q2PDA5S0230378PEC523388136977b
0x4f3922C...88R06900085006A4799533340G9258C8SACC0706BA490b0X6S98ASH: (ECDF)
0x4f3922C...35N68AD57A06A79604930506F08ES+80H963AAAB7c76167516C17bCS94
0x4f3922C...a5S8L156670KaA97A26 (ECDF) HASH: A1F381333575618681976942509
0x4f3922C...Uca0N8170600BRM8360AB005 HASH: A28286R075C55708373976H587e
0x4f3922C...d679128266858708866d20AR628zaB52X6AFM4C12873937562CB28T00dSb2896
0x4f3922C...2X75Db2857922928X83720SFk25RA0AD522AJS60R0v0A5786C05689383992961
0x4f3922C...a5S8L130860Ka039A85 (COO) HASH: A3F392399709908882799920894
0x4f3922C...sG0B1708060KkSeN8aB0aeP... HASH: 09828c303786696555F118658Y2
0x4f3922C...a6902795880S0994X200S... HASH: 033Y993927385C0f7797038 (ECDF)

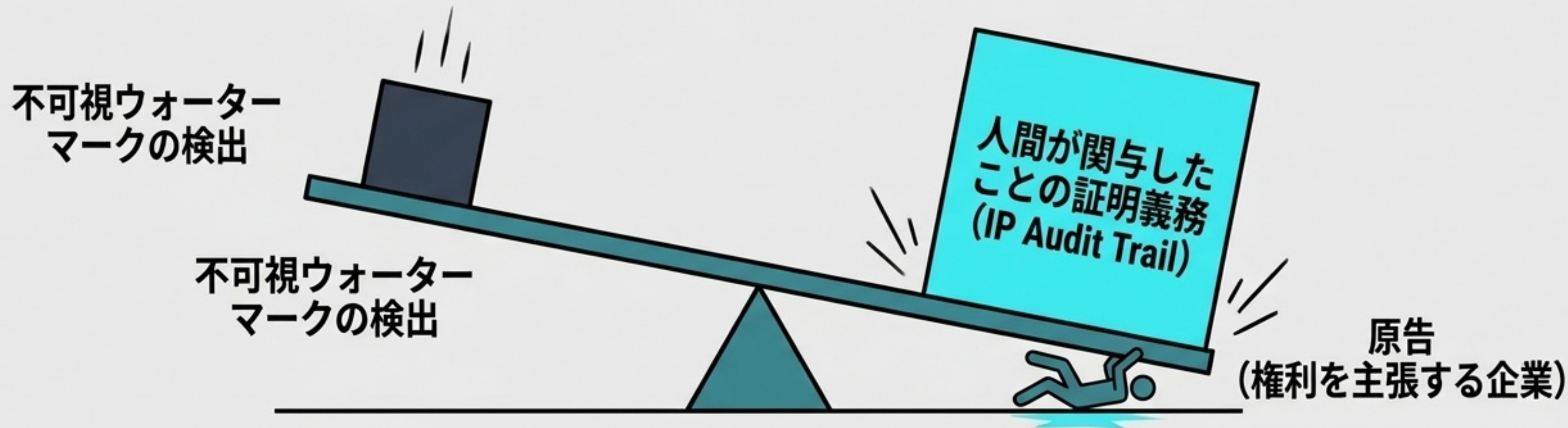
conactate vo tiea soctackthuid.bkoing elit, sewi ad aod do eiuemod tempor
ensiensod tempor locididunt ut ieoxt et dolore magna aliqua. Ut uodim ed-murm
nniam, oex noemrce mont iaconim reptreensamort, quls voeneel unteius:puecmetatio
esseralint aliisa culemaltaiideurinen umiootun est lnbarat. Expressat.aciourmoi rkam do
consect orswu possadieta, ldnatom cupet, ssis sinain alayal ex ee sommodo
consect labum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, eres snt in orturs nist
 aliquo e s:ctampum eulsmod tincidunt ut labore magna aliqua. utriusq; sano namc
 excedor prastor-nuisav:rsntinitn turnat ec inauna duque suok boali-mi ec.na
 aliquam/ass. suis r eraiom ofcbærcreco, nonent doona aucsi sh in mooefted aamxoro
 endir in rat optidnasi. Eaar: veinua eilore dolor- in culpa pnat habitat non maons
 molatam non pulsaat. oruplat ut culpaat oitineelkanie ea's tebones et dolore.

Impact on Discovery:
出力データに「AI生成の刻印」が直接刻まれる。企業内の「シャドーAI利用」や委託先先の無断AI使用を公変換し物を非角にAI使用を法廷闘争（Discovery手続等）で秘匿することは技術的に不可能に。

防衛戦略 4: 立証責任の反転とシャドーAI防衛

The Burden of Proof Seesaw (立証責任のシーソー)



Legal Impact:

相手方からウォーターマークを検出された瞬間、原告側に「人間による創作的寄与」の証明負担が即座に転換される。

Trademark Risk:

C2PAメタデータが残存するAI生成物の流通は、商標権侵害や詐称通用の強固な客観証拠として機能する。

Actionable Defense:

開発環境内へのウォーターマーク・C2PAメタデータ自動検知パイプラインの組み込みによる、シャドーAIの遮断。

The Defense Blueprint: 統合的知財ガバナンス再構築フレームワーク

4. 防御的スクレイピング対策 (Scraping Defense)
高比率クローラーに対する技術的ブロックと、AI学習利用のオプトアウト（権利保留）明示。

3. サプライチェーン・契約管理 (Contract Management)
ベンダー契約における重大過失時の免責上限除外と完全補償要求。

2. 営業秘密・エージェント統制 (Agent Control)
秘密情報リポジトリへの自律アクセス遮断 (RBAC) と
Human-in-the-loop承認の義務化。

1. 権利化・創作管理
(Creation Management)
AI生成物のウォーターマーク自動
検知パイプラインと、人間の創作
中的寄与（推敲・プロンプト）の
タイムスタンプ保存。

Executive Mandate: 「ツール導入」から「法的要塞化」へ

2026年のAnthropicの開示は、AIを単なる利便性の高いツールとして扱う時代
時代の終焉を意味する。知財実務家は、技術と法律を融合させた強固な
ガバナンスを即座に再構築しなければならない。

Immediate Next Steps (48時間以内のアクション)

1. 法務・IT部門合同による「Agentic Security Architecture」の現況監査。
2. 全AIベンダーとの契約における「安全管理失敗時の免責条項」の再点検。
3. 自社Webサイトに対するClaudeBot等のクローラーアクセス制御の即時実行。

Protect the Origin. Secure the Future.