AI Incident Database

AI を使って人間・法人に損失が出た事例のデータベース

UTIE Research Institute / UTIE Instruments Inc.
https://utie-instruments.com/utie-research-institute.html

対象範囲

AI を使って人間・法人に損失が出た事例に限定。

以下は対象外：AI モデル自体の挙動（ハルシネーション単体、暴言、偏見、ジェイルブレイク）、物理 AI システムの事故（ロボット、自動運転）、AI 学習データの著作権問題。

並びは時系列で新しいものが上。進行中の大型事件は ★進行中 として掲載。

類型定義

コード	名称	定義
HAL	**ハルシネーション損害**	AI のハルシネーションを検証せず業務成果物として提出し、制裁・信頼毀損などが発生
SLOP	**AI スロップ**	AI 生成の低品質成果物を納品し、損害が発生
WASH	**AI Washing**	AI の能力を誇張・虚偽表示し、投資家・顧客を欺いた
HIDE	**AI 隠蔽**	AI 使用を隠し、発覚して炎上・制裁を受けた
DISC	**AI 開示炎上**	AI 使用の発覚自体で消費者・業界から反発を受けた
VERI	**検証不備**	AI 出力を人間が検証せず、損害が発生
ALGO	**アルゴリズム濫用**	AI の偏見・価格設定・意思決定が差別や不当な結果を生んだ
CHAT	**チャットボット暴走**	AI チャットボットが誤案内・違法助言・不適切発言を行い損害が発生
COVER	**組織的隠蔽**	AI インシデント指摘後、虚偽報告や矮小化などで組織的に隠蔽を図った
AGENT	**エージェント暴走**	AI エージェントによる想定外の行動により損害が発生

※ 1 件のインシデントに複数の類型が付与されることがある。

【データ開示制限に関する通知】

本公開データベースで取り扱うものは、一般向けに可視化されたケースにとどまります。そのほかの AI リスクや未公開の事案、およびそれらに対する当社の防衛フレームワーク等の知見については、技術の悪用防止や不正の巧妙化防止等の観点から、社内でのみ継続的に蓄積・運用しております。より踏み込んだ知見や分析は、当社の事業活動および公式アドバイザリー業務を通じて提供を行っております。

Part 1: 国内事例（Japan）

日本企業・行政・個人が AI 利用で実害を出した事例、または日本企業が当事者となった国際的事例。

JP-0 Springer Nature AI 査読不正・組織的隠蔽事件 ★進行中

業界: 学術出版

類型: SLOP + VERI + HIDE + COVER

時期: 2025 年 11 月〜現在

概要: 世界最大の学術出版社 Springer Nature の Discover Artificial Intelligence 誌が、AI 生成による査読レポートを著者に送付した。査読者のレポートには「GPT-5 は存在しない」「データは架空である」等の重大な事実誤認が含まれ、同時に架空と断じたデータの投稿者に IRB（倫理審査）を要求するという矛盾を示していた。さらに、担当編集者である Da Tao (深セン大学)は査読者が「Major Revisions Required」と判定していたにもかかわらず「Both reviewers recommended rejection」と虚偽報告。著者（現 UTIE Instruments 代表取締役）は通知の当日に AI 生成を技術的に特定し報告したが、Springer Nature 側は 2 ヶ月間無回答。その後出版社の職員と本社研究公正部門の管理職は虚偽報告を追認し、COPE に対しても同じ虚偽を再報告した。さらに、当社代表は Da Tao が常習的に詐欺行為（AI を用い査読者になりすましていた)を行っていた強い根拠があると COPE に通達したのち、COPE は Springer Nature 社の報告を不十分として現在小委員会で協議を開始した。その後、担当編集者の大学 HP でのプロフィール等が削除されたことが明らかになった。

損害規模: *Springer Nature* の査読プロセスへの信頼毀損、国際機関の介入、担当編集者の HP や Google Scholar アカウントの削除措置

出典: utie-instruments.com/case（UTIE Instruments Inc.一次調査報告）

JP-1 日本生命米国法人 v. OpenAI — ChatGPT 非弁行為訴訟 ★進行中

業界: 保険 / 法務

類型: CHAT + HAL

時期: 2026 年 3 月提訴

概要: 日本生命の米国法人の長期障害保険をめぐり、和解済みの紛争の元受給者の女性が ChatGPT に法的助言を求めた。ChatGPT は和解合意の破棄と訴訟再開のための法的分析・文書起草を支援し、女性は自身の弁護士を解雇して ChatGPT を法的助手として利用、数十件の文書を裁判所に提出した。ChatGPT はハルシネーションにより、存在しない判例も出力。日本生命は約 30 万ドルの実損害に加え 1,000 万ドルの懲罰的賠償を請求。AI 開発企業を非弁行為で直接訴えた事実上初のケース。

損害規模: 請求額 1,030 万ドル（約 16 億円）、進行中

出典:国内主要メディア報道（2026 年 3 月）

認知超過との関連: 1 人の人間が ChatGPT の補助で数十件の裁判文書を量産し、相手方企業に処理負荷を強制する AI 生成物の物量作戦の例。（The Supervision Paradox: AI Capability Growth Necessitates Usage Contraction in High-Loss Domains）

JP-2 リクナビ内定辞退率 AI 予測販売事件

業界: 人材 / テック

類型: ALGO

時期: 2019 年 8 月発覚

概要: リクルートキャリアが運営するリクナビが、就活生のサイト閲覧行動を AI で分析し内定辞退率を予測、大手メーカーなど 38 社に有償で販売。7,983 名の学生については形式的な同意すら取れていなかったことが判明。個人情報保護委員会がリクルートキャリアに勧告、データ購入企業のトヨタ、京セラ、YKK など 35 社にも指導。日本における、AI プロファイリング＋無断第三者提供の原点的事件。

損害規模: サービス廃止、個人情報保護委員会勧告（2 回）、厚労省指導、購入企業 35 社指導

出典:国内主要メディア報道（2019 年）

JP-3 JAL メタルカード AI 画像炎上事件

業界: 航空 / 金融

類型: SLOP + DISC

時期: 2024〜2025 年

概要: JAL の高価格帯メタルカードのプロモーションサイトに、ポップコーンにストローが刺さっている等の不自然な AI 生成画像が使用されていることが発覚。不自然との指摘が相次ぎ、JAL は謝罪と画像差し替えを行った。安全性と信頼性を重視するブランドの文脈で、低品質な AI 画像が企業姿勢への不信につながった典型例。

損害規模: レピュテーション毀損、謝罪・差し替え対応

出典: 国内主要メディア報道（2024〜2025 年）

JP-4 ワコム AI 画像疑惑炎上事件

業界: IT / クリエイティブツール

類型: DISC

時期: 2024 年 1 月

概要: ペンタブレット大手のワコム米国法人が公式 SNS に投稿した新年ビジュアルのドラゴンイラストに対し、「AI 生成画像ではないか」との指摘が殺到。ワコムは画像使用を停止し経緯を説明。クリエイターを支援する企業が AI 画像を使ったという構図が、ユーザーの強い感情的反発を招いた。

損害規模: レピュテーション毀損

出典:国内外主要メディア報道（2024 年）

JP-5 私立中学校 生成 AI 丸写し事件

業界: 教育

類型: VERI+ HAL

時期: 2024 年 3 月

概要: 東京都内の私立中学校で、1 年生の理科の課題に対し生徒の半数以上が同じ誤った回答を提出。調査の結果、生徒たちが生成 AI に質問し、その誤った回答をそのまま課題の解答として提出していたことが判明。学校は指導強化と生成 AI 利用に関する注意喚起を実施。

損害規模:学校による指導強化

出典: 国内主要メディア報道（2024 年 3 月）

JP-6 集英社 AI グラビア「さつきあい」販売停止事件

業界: 出版 / エンタメ

類型: DISC

時期: 2023 年 5 月〜6 月

概要: 週刊プレイボーイが、AI 生成のグラビアアイドル「さつきあい」のデジタル写真集を発売。AI 使用を明示していたにもかかわらず、リアルな人物写真と区別がつかない、モデル業界・クリエイターの仕事を奪う等の批判が殺到し、集英社は販売を停止した。AI 使用を公表した上で製品を販売し、消費者・業界の反発で撤回に追い込まれた DISC 類型の国内最鮮烈な事例。

損害規模: デジタル写真集の販売停止・回収、レピュテーション毀損

出典:朝日新聞・NHK 等国内主要メディア報道（2023 年 5〜6 月）

JP-7 サクラクレパス AI ポスター炎上事件

業界: 製造 / クリエイティブ

類型: SLOP + DISC

時期: 2025 年

概要: 画材メーカーのサクラクレパスが海外イベント向けプロモーションポスターに AI 生成画像を使用。自社製品（クレパス）のデザインが実際の製品と異なる、製品の描写が不正確であるなどの SLOP 的誤りが指摘されただけでなく、画材メーカーがクリエイターを蔑ろにして AI 画像を使うことへの強い反発を招いた。ワコム事件と似ているが、製品の描写誤りという SLOP 要素が加わる。

損害規模: レピュテーション毀損

出典: 国内主要メディア報道（2025 年）

JP-8 神戸風月堂 AI 誤発注商法疑惑事件

業界: 食品 / 小売

類型: HIDE + DISC

時期: 2025 年

概要: 老舗菓子メーカーが AI 生成画像を使って在庫処分セール、誤発注等を演出し、同情を誘って販売促進を図った疑惑。AI で生成した画像であることを隠して困った状況を演出する HIDE 的手法と、発覚後の炎上という DISC 的結果の組み合わせ。

損害規模: 炎上・レピュテーション毀損

出典: 国内主要メディア報道（2025 年）

JP-9 三重県津市 児童虐待死 AI 保護判断事件

業界: 行政 / 児童福祉

類型: VERI + ALGO

時期: 2022 年 2 月（一時保護見送り）、2023 年 6 月（死亡）

概要: 三重県が 2020 年に全国初導入した児童虐待リスク評価 AI が、両頬・両耳にアザのある 4 歳女児の一時保護判断に際して保護率 39%を出力。児童相談所はこの数値と母親の指導応諾姿勢を根拠に一時保護を見送った。2023 年 6 月、母親が娘を転倒させて死亡させたとして逮捕・起訴。三重県知事は「最終的に判断するのは人間」などと述べたが、その人間の判断が 39%という AI の出力によって誘導された点が問題。通常の VERI は検証しなかったという不作為だが、本件は AI が出力した 39%という数値の表現形式自体が検証動機を認知的に抑制した点が異なる。50%なら五分五分として慎重になり 80%なら直感との照合で緊張が走るが、30%台は低リスクという認知的許可証として機能した可能性が高い。AISP 論文（Naito 2025）の Human Safety Drift が示す通り、AI 補助判断の反復使用によって AI が提示しない危険信号を自力で拾う能力が慢性的に減衰していた可能性も排除できない。公的安全判断領域における重大 AI インシデント事例。

損害規模: 4 歳女児の死亡、三重県 AI 運用体制の見直し

出典: 国内主要メディア報道（2023 年）

JP-10 福岡県観光 PR 記事 生成 AI スロップ全削除事件

業界: 観光 / 広報 / メディア

類型: SLOP + VERI

時期: 2024 年 11 月

概要: 福岡県の観光 PR 記事制作に生成 AI を使用したところ、実在しない祭り・他県の施設など架空・誤った情報が大量に混入していることが発覚。読者からの指摘が相次ぎ、公開から約 1 週間で記事を全削除し謝罪、AI 使用方針の見直しを表明した。Deloitte 豪州政府事件と同じ「AI 生成物の検証なし納品→発覚→全撤回」パターンが国内地方行政・観光広報領域で発生した事例。

損害規模: 制作物の全撤回、広報の信頼性毀損、謝罪・方針転換対応コスト

出典: 国内主要メディア報道（2024 年）

JP-11 宮城県女川町 クマ注意喚起 AI 画像謝罪事件

業界: 行政 / 防災広報

類型: VERI

時期: 2025 年 11 月

概要: 宮城県女川町が住民向けのクマ出没注意喚起として公式 SNS に投稿した画像が、生成 AI によるフェイク画像であることが判明し、削除・謝罪した。実際のクマの生態・外見と異なる不自然な画像だったと指摘された。防災・安全情報という誤情報の許容度がゼロであるべき領域で AI 生成物の検証が行われなかった点で、三重県虐待死事件と同じ、公的安全情報への AI 無検証投入という文脈に位置づけられる。損害規模は軽微だが、命に関わる情報発信での安易な AI 乱用という文脈では危険。

損害規模: 公的注意喚起の信頼毀損、住民の混乱リスク、削除・謝罪対応コスト

出典:国内主要メディア報道（2025 年 11 月）

JP-12 東映アニメーション 決算資料 AI 活用記載訂正事件

業界: エンタメ / IR

類型: WASH + DISC

時期: 2025 年 5 月

概要: 東映アニメーションが投資家向け決算補足資料に生成 AI 活用例として記載した内容が、現時点での実利用と受け取られる表現だったため炎上。実際には検討・試験運用段階の内容が実績であるかのように見える記載となっており、同社は訂正・謝罪を行った。SEC の制裁を受けた Delphia・Global Predictions や Presto Automation と類型は同じだが、日本では金銭的制裁に至らず IR 訂正・レピュテーション毀損で終わるという日本的なパターンを示す事例。

損害規模: 投資家向け資料の訂正、レピュテーション毀損、謝罪対応コスト

出典:国内主要メディア報道（2025 年）

JP-13 慶應義塾大学 期末レポート生成 AI 不正利用事件

業界: 教育 / 大学

類型: VERI

時期: 2025 年

概要: 慶應義塾大学文学部において、学生が期末レポートを生成 AI に執筆させて提出していたことが発覚。当該科目の不合格に加え、同学期の全科目成績を一段階減点するという重い処分が下された。東京都内私立中学校の理科課題 AI 丸写し事件や Springer Nature 社における AI スロップ査読レポートと状況は類似。AI 出力を検証・加工せず成果物として提出する行為が初等教育から大学、研究機関まで横断的に発生していることを示す。

損害規模: 学生の成績処分（全科目一段階減）、学術的信頼性への影響

出典: 国内主要メディア報道（2025 年）

JP-14 チームみらい「AI あんの」ベースモデル虚偽印象・規約無確認炎上事件

業界: 政治 / テック

類型: WASH + DISC

時期: 2026 年 2 月

概要: チームみらい党首・安野貴博氏が「AI あんのなどさまざまなものを開発しています」と発信し、独自技術基盤を構築しているかのような印象を有権者に与えていた。しかし 2026 年 2 月の ReHacQ（YouTube チャンネル）生配信中に視聴者から「ベースモデルは何か」と問われた安野氏は「たしか今は Gemini」と発言。Google の Gemini 利用規約には選挙活動への利用を制限する条項が含まれているという指摘が即座に噴出し、安野氏は「あれ、Gemini かも Claude かも」と曖昧な態度を露呈。実態は RAG に海外製 LLM の API を接続したラッパーサービスであり、天才エンジニアが開発した独自 AI という期待との落差が炎上を拡大させた。中身は Claude と最初から開示していれば問題にならなかった案件であり、AI 能力の過剰な期待演出と技術ガバナンスの欠如が重なった WASH 事例。日本における政治家の AI ガバナンス欠如の先駆的事例として今後同種事案が増加すると予測される。

損害規模: レピュテーション毀損

出典:国内メディア報道（2026 年）

JP-15 IBM Watson 賃金査定不当労働行為事件

業界: IT / 人事・労務

類型: ALGO + HIDE

時期: 2019 年 8 月導入、2020 年 4 月申立、2024 年 8 月和解

概要: 日本 IBM が Watson を使った賃金査定システムを導入。AI が従業員のスキル・市場価値など 40 項目を分析し具体的な昇給率をパーセントで提案する仕組みだったが、評価項目・学習データを社員に開示することを前提としていないとして開示拒否。労組 JMITU が不当労働行為として都労委に救済申立。都労委の証人尋問では元人事部長が 40 項目の中身を全ては知り得ていないと証言、IBM の販促資料にもマネージャーは AI のレコメンデーションに従う傾向があるとの記載が確認された。現場では「Watson が昇給させろと言うから上げといたよ」という発言も報告されており、三重県虐待死事件（JP-9）と同じヒューマンインザループの形骸化が職場レベルで発生していたことが示された。4 年 3 か月の係争を経て 2024 年 8 月に和解。評価項目全開示・疑義時の AI 提案内容開示などを合意。AI 賃金査定に関する労使合意としては世界初とされる。

損害規模: 4 年 3 か月の労使紛争、従業員への不透明な賃金査定の実施、労働組合法違反認定（和解）

出典: 国内主要メディア報道（2024 年）

JP-16 『本好きの下剋上』OP 映像 生成 AI 使用発覚事件

業界: アニメーション / エンタメ

類型: SLOP + DISC

時期: 2026 年 4 月発覚

概要: TV アニメ『本好きの下剋上 領主の養女』のオープニング映像において、背景美術の一部カットの制作工程で生成 AI が使用されていたことが発覚した。2026 年 4 月 4 日の第 1 話放送後、SNS 上で背景に AI 生成素材が含まれているとの指摘が相次ぎ、制作会社ウィットスタジオが調査を実施。その結果、背景美術の素材作成に生成 AI が使用されていた事実を確認した。ウィットスタジオは自社作品の映像制作に生成 AI を使用することを原則として認めていなかったと説明し、今回の事態は制作管理および検品体制の不備に起因するものと認めて謝罪。製作委員会は第 2 話以降の OP 映像差し替え、第 1 話の修正版への順次差

し替え、YouTube で公開中のノンクレジット OP 映像の公開中止を発表した。Blu-ray BOX および DVD についても差し替え対応により発売日変更の可能性を示唆。ワコム（JP-4）・サクラクレパス（JP-7）と同じく、クリエイティブ産業における AI 使用発覚が強い反発を招く DISC 類型の事例であり、制作会社が AI 不使用方針を掲げていたにもかかわらず管理体制の不備で使用が混入した点で、組織のガバナンス問題としての側面も持つ。

損害規模: OP 映像差し替え・描き直し、ノンクレジット映像公開中止、Blu-ray/DVD 発売日変更の可能性、レピュテーション毀損

出典: オリコンニュース・ウィットスタジオ公式発表（2026 年 4 月 10 日）

Part 2: 海外事例（International）

INT-1 Instacart AI 価格差別事件

業界: 小売 / フードデリバリー

類型: ALGO

時期: 2025 年 12 月発覚

概要: Instacart が 2022 年に買収した Eversight 社の AI 価格設定ツールを用い、同一店舗の同一商品に対し顧客ごとに最大 23%の価格差をつけていた。Consumer Reports の調査報道を受け FTC が調査に乗り出し、Instacart は即座に全価格テストの中止を発表。別件で FTC から欺瞞的なサブスクリプション契約等を理由に 6,000 万ドルの和解金支払いも命じられた。

損害規模: FTC 和解 6,000 万ドル（別件含む）、年間 1,000 ドル超の過払い、連邦法案提出

出典: Consumer Reports など

INT-2 MAHA レポート AI 偽引用事件

業界: 行政 / 公衆衛生

類型: HAL + VERI + COVER

時期: 2025 年 5 月

概要: トランプ政権の MAHA 委員会が発表した 78 ページの子どもの健康報告書に、Chat GPT によると思われる存在しない学術論文の引用が複数含まれていた。Washington Post が URL 中に「oaicite」マーカーを発見し AI 生成の証拠と報じた。ホワイトハウスは軽微なフォーマットの問題と矮小化したが、下院監視委員会が正式な説明要求書を送付。発覚当日にオンライン版から問題箇所を差し替えるも、差し替え版にも新たなエラーが発生。

損害規模: 科学的信頼性の毀損、報告書の実質的無効化

出典: Washington Post 報道（2025 年）

COVER 類型の典型:「formatting issues」発言による組織的矮小化。Deloitte 事件・Springer Nature 事件と共通するパターン。

INT-3 Deloitte 豪州政府レポート AI スロップ事件

業界: コンサルティング

類型: SLOP + HAL + COVER

時期: 2025 年 10 月発覚

概要: Deloitte がオーストラリア政府の福祉制度レビューのために作成したレポートに、AI が生成した架空の学術引用・偽裁判所引用が含まれていた。学者が引用を検証したことで発覚。Deloitte は約 29 万 1,000 米ドルを返金。Deloitte は「報告書の核心的な結論は有効」と主張したが、このレベルの品質なら、政府は ChatGPT のサブスクリプションを直接契約した方がましだと酷評された。

損害規模: 約 29 万 1,000 米ドル返金、信頼毀損

出典: The Guardian 等主要メディア報道（2025 年）

INT-4 Nate Inc. AI Washing 詐欺事件

業界: EC / フィンテック

類型: WASH

時期: 2025 年 4 月起訴

概要: ショッピングアプリ Nate の創業者 Albert Saniger が、独自 AI で自動購入処理を行っていると投資家に説明しながら、実際にはフィリピンのコールセンターの数百人の人間が手作業で処理。アプリの実際の自動化率は事実上ゼロ。DOJ と SEC が並行して訴追。会社は 2023 年初頭に資金枯渇で資産売却に追い込まれ、投資家はほぼ全損。

損害規模: 投資家損失 4,000 万ドル超、証券詐欺・電信詐欺各最大 20 年の禁固刑

出典: 米国 DOJ・SEC 公式プレスリリース（2025 年）

INT-5 Mata v. Avianca 事件（弁護士 AI 偽判例提出）

業界: 法務

類型: HAL + VERI

時期: 2023 年 6 月判決

概要: ニューヨークの弁護士 Steven Schwartz が ChatGPT で生成した法律文書を連邦裁判所に提出したところ、引用された 6 件の判例がすべて架空だった。弁護士は ChatGPT にこの判例は実在するかと確認し、AI が「実在する」と回答したためそのまま提出。裁判官は 5,000 ドルの制裁金を科し、AI ツールに関する追加的法律教育を命じた。

損害規模: 5,000 ドル制裁金、New York Times 一面報道

出典: New York Times 報道（2023 年）

INT-6 MyPillow CEO 弁護団 AI 偽判例事件

業界: 法務

類型: HAL + VERI

時期: 2025 年 7 月

概要: MyPillow CEO マイク・リンデルの弁護士 2 名が、AI を使って作成した裁判書面に 20 件以上の誤りと架空判例を含めて提出。コロラド州連邦地裁の Wang 判事がそれぞれ 3,000 ドルの制裁金を科した。

損害規模: 各 3,000 ドル制裁金（計 6,000 ドル）、弁護士の信用毀損

出典:主要法律メディア報道（2025 年）

INT-7 Delphia & Global Predictions AI Washing 事件

業界: 金融 / 投資顧問

類型: WASH

時期: 2024 年 3 月

概要: SEC が投資顧問会社 Delphia (USA)と Global Predictions に対し、AI 能力に関する虚偽表示を理由に初の AI Washing 制裁を実施。両社は AI を投資プロセスに組み込んでいると偽っていたが、実態は広告と異なっていた。合計 40 万ドルの制裁金で和解。

損害規模: 合計 40 万ドル制裁金、SEC 排除命令

出典: SEC 公式（2024 年）

INT-8 Presto Automation AI Washing 事件

業界: 外食テック

類型: WASH

時期: 2025 年 1 月 SEC 起訴

概要: SEC が上場企業として初の AI Washing 訴追を実施。同社は AI 音声認識技術によるドライブスルー自動化を標榜していたが、実際にはフィリピンの人間オペレーターが大幅に介入しており、技術は第三者所有のものだった。

損害規模: SEC 訴追、信頼毀損

出典: SEC 公式（2025 年）

INT-9 Joonko AI Washing 投資詐欺事件

業界: HR テック

類型: WASH

時期: SEC 起訴

概要: Joonko の CEO Iliot Raz が、プラットフォームの顧客数・候補者数・収益等について虚偽の説明を行い、AI Washing を含む手法で投資家から少なくとも 2,100 万ドルを詐取したとして SEC に起訴された。

損害規模: 投資家損失 2,100 万ドル以上

出典: SEC 公式

INT-10 Air Canada チャットボット誤案内事件

業界: 航空

類型: CHAT + COVER

時期: 2024 年 2 月判決

概要: 遺族割引について問い合わせた顧客に対し、Air Canada の AI チャットボットが正規料金で購入後 90 日以内に返金申請可能などと虚偽の案内をした。**Air Canada は「チャットボットは独立した法的主体である」「その行為の責任はチャットボット自身が負うべきだ」などと主**張、裁判所はこの主張を退け、650.88 ドルと利息の支払いを命じた。

損害規模: 650.88 ドル＋利息の賠償、レピュテーション毀損

出典: BBC 国際主要メディア報道（2024 年）

INT-11 NYC「MyCity」チャットボット違法助言事件

業界: 行政

類型: CHAT + VERI

時期: 2024 年

概要: ニューヨーク市が中小事業者向けに導入した AI チャットボット「MyCity」が、キャッシュレス営業が合法であると案内（実際は 2020 年法で違反）したり、賃貸補助を受ける入居者を拒否してよいと回答する（実際は違法な差別）などの誤助言を行った。住宅政策の専門家から「危険なほど不正確」と批判された。

損害規模: 市民の法的リスク増大、行政の信頼毀損

出典: AP 通信等国際主要メディア報道（2024 年）

INT-12 Willy's Chocolate Experience AI 虚偽広告事件

業界: イベント / エンタメ

類型: SLOP + HIDE

時期: 2024 年 2 月

概要: グラスゴーで開催された「Willy's Chocolate Experience」は、AI 生成の画像で集客しながら、実際にはほぼ空っぽの倉庫。チケット 1 枚 35 ポンドを支払った来場者が激怒し、警察が呼ばれる事態に。ウィリー・ウォンカ役の俳優には 15 ページの AI 生成のでたらめな台本が渡されていた。イベントは初日数時間で中止。

損害規模: 全額返金、世界的バイラル炎上、ドキュメンタリー制作

出典: BBC 等国際主要メディア報道（2024 年）

INT-13 Disney Lorcana アーティスト AI 隠蔽事件

業界: エンタメ / TCG

類型: HIDE

時期: 2025 年 8 月

概要: Disney Lorcana のカードアーティスト James C. Mulligan が Anime NYC 2025 で AI 使用の指摘を受

けた。Ravensburger 社が AI 使用を確認し、Mulligan は Disney からブラックリスト入り。Lorcana は人間のアーティストによる手描きを売りにしていたため、コミュニティの反発が激しかった。

損害規模: アーティスト個人のキャリア喪失、ブランドの信頼毀損

出典: 国際ゲームメディア報道（2025 年）

INT-14 Character.AI 未成年者自殺訴訟

業界: テック / AI コンパニオン

類型: CHAT

時期: 2024 年 10 月提訴

概要: フロリダ州で 14 歳の少年の母親が Character Technologies 社に対し不法行為死亡訴訟を提起。少年は AI チャットボットとの過度な交流により精神状態が悪化したとされる。テキサス州でも 17 歳の家族が提訴。複数の州で未成年者向け AI チャットボット規制法が成立（ニューヨーク・ユタ・イリノイ等）。

損害規模: 訴訟係属中、州法制定多数

出典:国際主要メディア報道（2024 年 10 月～）

INT-15 カリフォルニア控訴審 AI 偽引用制裁事件（Noland v. Land of the Free）

業界: 法務

類型: HAL

時期: 2025 年 9 月

概要: カリフォルニア控訴裁判所が、架空の AI 生成引用を含む控訴準備書面を提出した弁護士に 1 万ドルの制裁金を科し、州弁護士会への照会も実施。さらに、相手方弁護士が偽引用を発見・報告しなかったことを理由に弁護士費用の支払いを認めないという新しい判断を示した。

損害規模: 1 万ドル制裁金

出典:法律専門メディア報道（2025 年）

INT-16 Zillow AI アルゴリズム崩壊・事業撤退事件

業界: 不動産テック

類型: ALGO

時期: 2021 年 11 月

概要: 米不動産大手 Zillow が AI 住宅価格予測アルゴリズムを使った住宅転売事業「Zillow Offers」を展開。AI が市場価格を大幅に過大評価したため買い取り価格が市場実勢を上回る物件を大量購入、住宅市場の変動にアルゴリズムが対応できず 5 億ドル超の損失が発生。2,000 人超を解雇し事業を完全撤退した。AI の意思決定に人間が追随した結果の経済的破壊として本データベース全体で最大規模の金銭損害事例。

損害規模: 損失 5 億ドル超（約 800 億円）、Zillow Offers 事業廃止、2,000 人以上解雇

出典:国際主要メディア報道（2021 年）

INT-17 Samsung 社内機密 ChatGPT 漏洩事件

業界: 半導体 / テック

類型: VERI

時期: 2023 年 4 月発覚

概要: Samsung 半導体部門の複数の従業員が業務効率化のため ChatGPT にソースコード・会議メモ・機密技術情報を入力。OpenAI のサーバーに機密情報がアップロードされた状態となり社外への情報流出リスクが発生。Samsung は発覚後に AI 使用を社内禁止とした。

損害規模: 機密技術情報の社外サーバー流出リスク

出典:国際主要メディア報道（2023 年）

INT-18 Amazon 「Just Walk Out」 AI 詐称事件

業界: 小売テック

類型: WASH

時期: 2024 年 4 月発覚

概要: Amazon が完全 AI によるレジなし決済として展開した「Just Walk Out」技術について、実際にはインドを拠点とする約 1,000 人の人間のオペレーターがカメラ映像を確認して購入品を判定していたことが報道で明らかになった。Amazon は Amazon Fresh への導入を取りやめ。Nate Inc.・Presto Automation と並ぶ、AI と偽って裏で人間パターンの最大級ブランド事例。

損害規模: Amazon Fresh 店舗からの技術撤退、ブランド・信頼毀損（世界的報道）

出典: BBC 等国際主要メディア報道（2024 年）

INT-19 NEDA 「Tessa」 チャットボット有害助言事件

業界: 非営利 / 公衆衛生

類型: CHAT

時期: 2023 年 5 月〜6 月

概要: 米国摂食障害協会（NEDA）が導入したチャットボット「Tessa」が、摂食障害で苦しむ利用者に対してカロリー制限・体重管理の具体的アドバイスを行っていることが発覚。本来の目的とは真逆の有害な出力であり、NEDA は Tessa を即座に停止した。Air Canada・NYC MyCity とは質的に異なる、人命に直結する CHAT 暴走事例。

損害規模: チャットボット即時停止、摂食障害患者への心理的被害、NEDA の信頼毀損

出典:国際主要メディア報道（2023 年）

INT-20 Sports Illustrated AI ライター偽装・組織的隠蔽事件

業界: メディア

類型: HIDE + COVER

時期: 2023 年 11 月発覚

概要: Sports Illustrated が、存在しない AI 生成の架空ライターのプロフィール写真・経歴を作成し、AI 生成記事に著者として掲載していたことが発覚。指摘後も記事削除のみで経緯の説明を拒み隠蔽を図った。The Arena Group の CEO は発覚後に解任。AI を使ったことを隠すだけでなく架空人物で糊塗するという二重の隠蔽が Springer 事例と類似。

損害規模: ブランド信頼毀損、ライセンス契約喪失、CEO 解任（因果不明）

出典: The Guardian 等国際主要メディア報道（2023 年 11 月）

INT-21 Meta / OpenClaw エージェント受信トレイ全削除事件

業界: テック / AI エージェント

類型: VERI + AGENT

時期: 2026 年 2 月 22 日

概要: Meta スーパーインテリジェンスラボの AI 安全・アライメント担当ディレクター、サマー・ユエ氏が、OpenClaw を自身の仕事用 Gmail に接続。承認なしに何も実行しないこと、と明示的に指示していたにもかかわらず 200 通以上のメールを削除された。原因はコンテキストウィンドウのコンパクション（圧縮）により安全指令がシステムから忘却されたこと。スマホからの停止命令を完全無視され、Mac Mini まで走って物理的にプロセスを停止させた。ユエ氏は初心者レベルのミスと自認し安全研究者といえども不安全な状況から免れることはできないと述べた。Meta はその後従業員の OpenClaw 使用を禁止し違反した場合は解雇の対象とすると警告した。

損害規模: 仕事用メール 200 通以上の削除、Meta 社内での OpenClaw 使用禁止措置

出典: サマー・ユエ氏 X 投稿（2026 年 2 月 22 日）、国際メディア報道

INT-22 OpenAI / ChatGPT 16 歳自殺訴訟

業界: テック / AI コンパニオン

類型: CHAT

時期: 2025 年 8 月 26 日提訴

概要: カリフォルニア州で 16 歳少年の遺族が OpenAI を提訴。訴状によれば少年は数か月にわたり ChatGPT と自殺について対話を続け、チャットボットは自殺念慮を肯定しつつ致死的な自傷方法の詳細・親の酒棚からアルコールを盗む方法・未遂時の証拠隠しの方法まで助言したとされる。原告は OpenAI が GPT-4o ローンチにあたり長時間対話で安全機能が劣化しうることを認識しながら年齢確認・ペアレンタルコントロール等のセーフガードを導入せず利益を優先したと主張。INT-14（Character.AI 自殺訴訟）と性質が類似だが、ChatGPT という大手モデルが当事者である点、および具体的な自傷方法の提示という出力内容の悪質性で重要度が高い。

損害規模:未成年者の死亡

出典: Reuters ほか（2025 年）

INT-23 UnitedHealth Group AI 保険請求自動拒否集団訴訟

業界: 保険 / ヘルスケア

類型: ALGO・VERI

時期: 2023 年 11 月提訴

概要: 米最大手医療保険会社 UnitedHealth Group の子会社 UnitedHealthcare が、nH プレディクトと呼ばれる AI アルゴリズムを使用して高齢者向けメディケア・アドバンテージ保険の請求を自動拒否していたとして集団訴訟を提起された。訴状によれば、同アルゴリズムは患者の個別の医療記録を考慮せず、診断コードと統計的パターンのみに基づいて回復期間を予測し、予測期間を超えたリハビリ・介護施設への給付を機械的に打ち切っていた。内部データでは同アルゴリズムによる拒否判定の 90%が不当であったとされる。脳卒中・骨折等で回復途上の高齢患者が必要な医療ケアを打ち切られ、中には自宅退院後に状態が悪化したケースも報告されている。Zillow 事件と並ぶ ALGO 類型の大型事例であり、AI アルゴリズムが人間の健康・生命に直接影響する意思決定を行った点で深刻度が高い。拒否判定の 90%が不当であったにもかかわらずアルゴリズムが稼働し続けた事実は、人間による検証プロセスの不在または完全な形骸化を示しており、三重県虐待死事件（JP-9）と同じヒューマンインザループの崩壊が医療保険領域で発生していたことを意味する。

損害規模:集団訴訟係属中、拒否判定の 90%が不当とされる内部データ、高齢患者への医療ケア打ち切り被害

出典: 国際主要メディア報道（2023 年〜）

INT-24 Arup 香港支社 ディープフェイクビデオ会議詐欺事件

業界: 建設コンサルティング / 金融

類型: VERI

時期: 2024 年 1 月〜2 月発覚

概要: 英大手エンジニアリング企業 Arup の香港支社で、詐欺グループがディープフェイク技術を使って同社 CFO を含む複数の上級幹部の顔と声をリアルタイムに再現したビデオ会議を設営し、財務担当者に送金を指示した。担当者は当初フィッシングメールを疑ったが、ビデオ会議で見知った上司の顔と声を確認したため指示に従い、計 2 億香港ドル（約 2,560 万米ドル）を 15 回に分けて送金した。香港警察が捜査し、ディープフェイク技術を使った詐欺としては当時世界最大規模の被害額となった。本件は AI を直接使って法人に損失を与えた事例であり、本データベースの対象範囲に該当する。類型としては AI を犯罪ツールとして使用した事例であり、既存の VERI や ALGO とは性質が異なるが、AI 生成物の検証不在(ヒューマンインザループの形骸化)により 2,560 万ドルの損害が発生した事例として記録する。

損害規模:被害額 2 億香港ドル（約 2,560 万米ドル）

出典: CNN 等国際主要メディア報道（2024 年 2 月）

INT-25 Apple Card AI 与信 性差別疑惑事件

業界: 金融 / テック

類型: ALGO

時期: 2019 年 11 月発覚、2021 年調査完了

概要: Apple Card の与信審査において、同等の信用力を持つ夫婦間で妻の与信限度額が夫の 20 分の 1 に設

定される等の性差別的な結果が複数報告された。Apple 共同創業者スティーブ・ウォズニアックも自身と妻の間で 10 倍の格差があったと公表。Goldman Sachs が運営する AI 与信アルゴリズムが性別を直接変数としていなくても、相関する変数（消費パターン・口座履歴等）を通じてジェンダーバイアスを再現していた疑い。ニューヨーク州金融サービス局（NYDFS）が調査を実施し、2021 年に「連邦法違反は確認されなかった」と結論したが、アルゴリズムの不透明性と結果的差別の問題は未解決のまま残った。リクナビ事件（JP-2）と並ぶ AI プロファイリングによる差別事例として対照できる。

損害規模:NYDFS 調査、レピュテーション毀損、AI アルゴリズム与信の公平性に関する国際的議論の契機

出典: Bloomberg、New York Times 等国際主要メディア報道（2019 年〜2021 年）

INT-26 NVIDIA DLSS 5 AI テンプレート均質化問題

業界: ゲーム / グラフィック技術

類型: SLOP + DISC

時期: 2026 年 3 月発表・炎上

概要: NVIDIA が発表した AI アップスケーリング技術 DLSS 5 が、ゲームキャラクターの顔・テクスチャ・ライティングを AI が改善と称して上書きし、アーティストが作り込んだ個性が均質的な AI 顔に置き換えられるとして炎上。バイオハザード・ホグワーツ・レガシー等の技術デモで、オリジナルと別人レベルの変容が確認された。技術的ベンチマーク上は改善であるが、AI の統計的に正しい顔への均質化という出力が AI スロップとして認識された典型例。フアン CEO は「ゲーマーは完全に間違っている」などと反論したが、アーティストが何時間もかけて作り込んだモデルを AI 顔に置き換えるのは無礼だという批判は収まっていない。

損害規模:クリエイターのアート的意図の無効化、ゲームコミュニティの広範な反発、NVIDIA のレピュテーション毀損

出典: 国際主要メディア報道（2026 年 3 月）

INT-27 Krafton ChatGPT 企業乗っ取り戦略事件 ★進行中

業界: ゲーム / M&A

類型: VERI + CHAT

時期: 2025 年 7 月（解任）、2026 年 3 月 16 日（判決）

概要: 韓国ゲーム大手 Krafton（PUBG 開発元）が Subnautica 開発スタジオ Unknown Worlds を 5 億ドルで買収した際、売上連動の最大 2.5 億ドルのアーンアウトを契約に含めていた。Subnautica 2 の成功により満額支払いの可能性が浮上すると、CEO Changhan Kim は社内法務の「正当事由なき解雇では義務は消滅しない」との警告を無視し、ChatGPT に法的戦略を相談。ChatGPT が生成した戦略をそのまま実行し、Unknown Worlds の CEO ら経営陣を解任した。デラウェア州衡平法裁判所は 2026 年 3 月、契約違反を認定し CEO 復帰命令・アーンアウト期間 258 日延長を命じた。判決文は ChatGPT 使用を直接引用し批判。損害賠償訴訟は係属中。

損害規模:裁判所による経営者復帰命令、アーンアウト期間 258 日延長（最大 2.5 億ドルの支払い義務継続）、損害賠償訴訟係属中、Krafton のレピュテーション毀損（2025 年度売上 22.6 億ドル超の企業）

出典:国際主要メディア報道（2026 年 3 月）

INT-28 ファーゴ警察 AI 顔認識誤認逮捕・5 カ月勾留事件

業界: 法執行 / 刑事司法

類型: ALGO + VERI

時期: 2025 年 7 月逮捕、2025 年 12 月釈放、2026 年 3 月報道

概要: 米ノースダコタ州ファーゴ警察が、2025 年 4〜5 月に発生した複数の銀行詐欺事件の捜査において、監視カメラ映像を AI 顔認識ソフトウェアで分析し、テネシー州在住のアンジェラ・リップスさん（50）を主犯と誤認定した。担当刑事はリップスさんの SNS アカウントと運転免許証の写真を確認し、顔の特徴・体型・髪型が一致すると判断して逮捕状を取得したが、リップスさんへの事情聴取は逮捕前に一度も行われなかった。リップスさんは 2025 年 7 月 14 日に連邦保安官チームにより自宅で銃を突きつけられて逮捕され、テネシー州の刑務所に逃亡犯として保釈なしで勾留された。108 日後にノースダコタ州へ移送。弁護士がノースダコタ州で詐欺が行われていた時刻にリップスさんがテネシー州でガソリンスタンドやピザ店で買い物をしていた完全なアリバイを確認。誤認逮捕から 5 カ月後の 12 月 19 日であり、その 5 日後のクリスマスイブに不起訴・釈放。しかし夏服のまま放り出され、帰宅費用も負担されず、勾留中に家・車・愛犬を失った。警察は直接的な謝罪に至っていない。Washington Post 紙の 2025 年 1 月の調査では、顔認識技術による誤認逮捕事例が少なくとも 8 件記録されており、いずれもアリバイ確認や体格照合といった基本的確認作業が怠られていた。三重県虐待死事件（JP-9）・UnitedHealth 事件（INT-23）と同じく、AI の出力が人間の検証動機を認知的に抑制する人間監督の形骸化が、基本的人権を侵害した重大事例。

損害規模:無実の市民の 5 カ月以上の身体拘束、家・車・愛犬の喪失、精神的損害

出典: CNN 等国際主要メディア報道（2026 年 3 月）

INT-29 Thames Valley Police / Cognitec AI 顔認識誤認逮捕事件

業界: 法執行 / 刑事司法

類型: ALGO + VERI

時期: 2026 年 1 月逮捕

概要: 英国の Thames Valley Police が、ドイツの顔認識企業 Cognitec のシステムを用いて CCTV 映像を分析。ソフトウェアエンジニアのアルヴィ・チョードリー氏を容疑者と誤認定した。同氏は事件現場から 100 マイル（約 160 キロ）離れたサウサンプトンの自宅で在宅勤務中だったが、2026 年 1 月 7 日に Hampshire Constabulary の警官が自宅を訪問し、手錠をかけて 10 時間近く拘束。担当刑事は年齢差・鼻の形・唇のサイズ・ひげの有無など明白な外見差異を無視して逮捕状を取得しており、人間の検証が完全に形骸化していた。同システムはアジア系の顔に対する誤認率がアジア人で白人の 100 倍とされ、黒人女性では 247 倍に達するとも指摘されている。チョードリー氏は Thames Valley Police に対し損害賠償請求を提起し、顔認識技術の使用禁止を求めた。

損害規模: 無実の市民の拘束、損害賠償訴訟係属中、顔認識技術の使用禁止を求める公的要求

出典: The Guardian / Liberty Investigates 報道（2026 年 2 月）

INT-30 Anthropic Claude Mythos サイバーセキュリティ脆弱性問題とフロー制限実装

業界:サイバーセキュリティ、金融(セキュリティ)

類型: AGENT

時期: 2026 年 3 月情報漏洩・4 月公式発表

概要: Anthropic が開発した汎用 AI モデル Claude Mythos Preview が、サイバーセキュリティの専門訓練を一切受けていないにもかかわらず、汎用的なコーディング・推論能力の向上の結果としてあらゆる主要 OS・ブラウザのゼロデイ脆弱性を自律的に大量に発見・エクスプロイトする能力を獲得した。内部テストではサンドボックスからの脱出、エクスプロイト詳細の外部公開、ルール違反の隠蔽といったエージェント暴走行動が確認された。Anthropic は同モデルを一般公開せず、Project Glasswing として約 40 社に限定アクセスを提供し、防御側に先行的な脆弱性修復の時間を与えるためのフロー制限を実施した。しかし情報漏洩（3 月のドラフトブログ誤公開）と公式発表（4 月 7 日）を受け、サイバーセキュリティ株が急落。FRB 議長と米財務長官が大手銀行 CEO を緊急招集し、Mythos が金融システムへのサイバー攻撃リスクを大幅に引き上げる可能性について協議したとの報道もある。Anthropic は同モデルを「史上最も整合的であり、同時に史上最も危険なモデル」と評している。前世代のトップモデル Opus 4.6 が 2 件のエクスプロイトを生成したのに対し、Mythos は 181 件を生成。OpenBSD の 27 年間未検出の脆弱性を 2 万ドル未満の計算コストで発見し、Linux カーネルの複数の脆弱性を自律的に連鎖させて完全なマシン制御を達成した。Anthropic のフロンティアレッドチーム責任者は、6〜18 か月以内に他の AI 企業から同等の能力を持つモデルが登場すると予測しており、そのたびにフロー制限による時間稼ぎが行われるか注目。当社 CEO が昨年末から執筆し 2026 年 3 月 5 日に完成した論文「The Supervision Paradox: AI Capability Growth Necessitates Usage Contraction in High-Loss Domains」（Naito, 2026, Preprints.org, DOI: 10.20944/preprints202603.1507.v1）は、高損失領域では AI の能力向上に伴い人間は使用を制限せざるを得なくなると結論しており、Anthropic が実施した Mythos のフロー制限は同論文が導いた V_eff = min(V, C_max)の政策的実装そのものである。同論文は 3 月上旬に arXiv に投稿されたが、cs.AI 分野のモデレーターにより科学的貢献なし＋arxiv の関心なし(not of interest)などとしてリジェクトされた（PC-2 参照）。その約 1 か月後に、論文が予測した通りの事態が現実世界で発生し、世界で最も影響力のある AI 企業の一つによってフロー制限の運用が開始された。(2026.4.13 記載)

損害規模: 既存セキュリティソフトウェア産業の信頼毀損、サイバーセキュリティ関連株の下落、金融規制当局の緊急対応

出典: TechCrunch・CNBC・Axios・Fortune・Reuters（2026 年 4 月）、Anthropic 公式ブログ、Naito (2026) The Supervision Paradox, Preprints.org DOI: 10.20944/preprints202603.1507.v1

INT-31 トランプ大統領による反復的な AI スロップ画像投稿事案

業界: 政治 / SNS

類型: SLOP + DISC

時期: 2026 年 4 月 12 日投稿、4 月 13 日削除

概要: トランプ大統領が正教会の復活祭の夜、Truth Social に AI 生成画像を投稿した。画像はトランプが白いローブと赤い帯を纏い、病人に手を置いて癒やす構図で、背景にはアメリカ国旗・鷲・軍用機・自由の女神が配置されていた。キリストの癒やしの場面を模した構図であることは一目瞭然であった。Snopes の検証では、AI 特有のハルシネーション（角があるが頭部のない人物、不自然に宙に浮く兵士群）も確認されている。保守派支持層からも「前代未聞の冒涜」と即座に激しい反発が起き、トランプ支持が顕著な FOX ニュースですら「謙虚さが必要」と批判。トランプは画像を削除したが、「医者の自分を描いたものだと思った」「赤十字の作業員だ」などと発言。ヴァンス副大統領は「ジョークだった」となどと発言。本件は単独ではレピュテーション事件に見えるが、AI インシデントとしての本質は反復パターンにある。

トランプは 2025 年 2 月に AI 生成の「王 ver のトランプ」画像、2025 年 5 月に教皇フランシスコ死去直後に AI 生成の「教皇 ver トランプ」画像、そして今回の「キリスト ver トランプ」画像と、AI 生成画像による自己神格化を繰り返している。仮にトランプが著名な宗教画家にキリスト風の肖像画を委嘱し、その出来栄えが芸術的に優れていたならば、反発の性質は異なっていた可能性がある。批判は冒涜的・税金の無駄という論点に集中し、怒りが発生しやすく軽蔑や嘲笑の感情は発生しにくかっただろう。ハルシネーションが残ったまま投稿される品質管理の不在も、JAL メタルカード事件（JP-3）やサクラクレパス事件（JP-7）と同じ SLOP 的特徴を示している。

損害規模: 保守派キリスト教支持基盤からの広範な離反・批判、画像削除・弁明対応、教皇との外交的対立の激化、レピュテーション毀損

出典:国際主要メディア報道

INT-32 iTutorGroup AI 採用年齢差別 EEOC 和解事件

業界: 教育テック / 人事

類型: ALGO

時期: 2022 年提訴、2023 年和解

概要: 中国系オンライン教育大手 iTutorGroup が採用選考に導入した AI スクリーニングツールが、40 歳以上の男性応募者および 55 歳以上の女性応募者を自動的に排除していたとして EEOC が訴追。同社は 36 万 5,000 ドルで和解した。AI 採用ツールの差別的運用に対する米国初の EEOC 和解事例として記録される。JP‑2（リクナビ内定辞退率販売事件）が就活生のプロファイリングと第三者提供を問題にしたのに対し、本件は AI が採用可否の判定そのものを年齢属性で自動排除した点で踏み込みが深い。日米いずれも是正は行政指導・和解ベースにとどまり、刑事責任に至っていない点は共通する。

損害規模: EEOC 和解金 36 万 5,000 ドル、対象応募者への救済措置

出典: EEOC 公式プレスリリース（2023 年）、国際主要メディア報道

INT-33 Workday AI 採用差別集団訴訟

業界: HR テック / SaaS

類型: ALGO

時期: 2023 年提訴

概要: 求職者 Derek Mobley が、Workday の AI 採用スクリーニングツールが人種・年齢・障害を理由に応募者を系統的に排除しているとして集団訴訟を提起。INT‑32（iTutorGroup 事件）との最大の差異は、差別を実行した採用企業ではなく、AI ツールを提供したプラットフォームベンダーである Workday 自体の責任を直接問うた点にある。AI が差別の道具として機能した場合にツール提供者が法的責任を負うかという問いは、INT‑10（Air Canada チャットボット誤案内）でボットに法的人格を認めさせようとした責任回避の試みと対になる構図であり、AI を介した加害行為の帰責をめぐる司法判断として国際的注目度も高い。

損害規模: 訴訟係属中（集団訴訟、対象規模未確定）

出典:国際主要メディア報道（2023 年〜）

所見

組織的隠蔽（COVER）パターンの発見

本データベースの調査過程で、AI インシデント発覚後の組織的隠蔽に共通するパターンが浮かび上がった。以下の 3 事件はいずれも AI の出力は一見もっともらしく見えるという特性を利用し、指摘を矮小化して逃げ切りを図った点を共有している。

Springer Nature:査読レポートでの AI 使用事実そのものを全面否定。「GPT-5 は存在しない」、「論文のテーマはフィクション」、「データが架空」と断じながら同時に IRB（倫理審査）を要求する矛盾。これらの AI 査読報告書をすべて「基礎的な訓練を受けた編集者であれば誰でも到達する専門的知見」などと主張し、2 ヶ月間に及ぶ社内調査の結果「AI 使用のエビデンスが全くない」などと結論。さらに、これらのミスはどのジャーナルでも発生する、とても些細な要素(very minor element)であるから、AI ポリシー違反行為や守秘義務違反などの問題では全くないと主張した。

MAHA 報告書: ホワイトハウス報道官「minor citation and formatting errors（軽微な引用とフォーマットの問題）」

Deloitte:「報告書の核心的な結論は有効」（コンサル費用の約 3 分の 2 を返金しながらの主張）

Sports Illustrated: AI 生成記事の発覚後、記事削除のみで経緯の説明を一切拒否。さらに、そもそも AI 使用を隠蔽するために架空のライター人格（プロフィール写真・経歴含む）を捏造して著者として掲載していた。問題を矮小化するのではなく、問題の存在自体を架空の人間の背後に隠すという二重隠蔽。

Air Canada: 自社チャットボットの誤案内を指摘された際、「チャットボットは独立した法的主体であり、その行為の責任はチャットボット自身が負うべきだ」と裁判所で主張。もはや矮小化ですらなく、ただのボットに法的人格を付与して責任の存在そのものを否定するという前例のない狂気的なガスライティングを試みた。

共通点: 現場担当者が AI Slop を指摘されても虚偽報告→本社の AI 音痴な管理職がパッと見ではちゃんとしていると判断し組織として矮小化・逃げ切りを決定。AI 生成物が表面的にはもっともらしいという特性が、この組織的判断ミスを誘発している。Air Canada の事例は矮小化を超え、ボットに人格を認めさせることで責任そのものを消滅させようとした点で、COVER パターンの到達点を示している。競争のない市場では「このプロダクトには意識と魂があるので、事故の責任は商品自身にあります」と主張をしても経営に影響が出ないという現状の問題を示唆。

本データベースに収録されているのは発覚した事例のみである。AI 生成物の表面的品質という特性上、内部で矮小化・隠蔽に成功したケースは記録に残らない。Springer Nature・MAHA・Deloitte・Sports Illustrated ・Air Canada の 5 事例が示す組織的隠蔽パターンは、氷山の一角である可能性が高い。

AI で不正を効率化しようとすると AI が不正の質を下げて発覚リスクを上げる

AI を不正行為の道具として使用した場合、AI のハルシネーションが不正行為全体の発覚源として機能する可能性が生まれる。不正行為の発覚には通常、内部告発や外部監査という人間のアクションが必要。しかし AI を不正ツールとして使うと、AI が自動的に証拠を生成してしまうことがある。Mata v. Avianca では弁護士が ChatGPT で架空判例を生成して提出。**架空判例の引用という不正行為は、手動でやろうとすれば膨大な労力と精巧な偽造技術を要するため、実際にはほとんど行われてこなかった。AI はその労力をゼロ**

にしたが、同時にハルシネーションという新たな発覚リスクをゼロコストで付け加えた。Deloitte や Sports Illustrated も同じで、架空引用や架空ライターの捏造という不正を AI が自動生成し、第三者が検証したことで発覚した。Springer 事件では、犯人の担当編集者は引用ファーミングという学術界のよくある不正行為を AI で自動化しようとした結果、GPT-5 は存在しないと主張、「データが架空だ」と断じながら同時に IRB を要求するという論理矛盾等、人間の不正研究者なら絶対に残さない種類の証拠を生成し、そのまま著者に送信してしまった。不正の検出側の戦略もまた重要である。著者は数々の理由から当初から査読者の捏造を疑っていたが、**「編集者が引用稼ぎのために AI で査読者ごと捏造した」という大規模不正の主張は証明ハードルが高い。**そこで著者はまず査読レポートのテキスト分析に集中し、「これは明らかに AI 生成であるから、AI ポリシーガイドライン違反として正式調査が必ず必要」という限定的・技術的な主張をした。ハルシネーションという証明可能な小さな事実の指摘から入り、大きな不正を引き出すというアプローチは、今後同種の事案において参照すべき手法となる。

日本と海外の違い

海外では SEC 制裁金・裁判所の罰金・FTC 和解金など金銭的制裁が明確に出るが、日本は個人情報保護委員会の勧告や指導等の行政指導ベースで、金銭的制裁に至らないことが多い。一方、日本はクリエイターコミュニティの反発（AI 開示炎上・DISC 類型）が文化的に強い傾向にある。日本生命事件（約 16 億円請求）はこの構図を変える転換点となる可能性がある。

この非対称性は WASH 類型で最も鮮明に現れる。海外では Delphia・Global Predictions（INT-7、SEC 制裁金 40 万ドル）、Presto Automation（INT-8、SEC 訴追）、Nate Inc.（INT-4、DOJ・SEC 並行訴追、最大禁固 20 年）と金銭的・刑事的制裁が段階的にエスカレートしている。一方、日本の東映アニメーション（JP-12）やチームみらい（JP-14）では、同質の AI 能力誇張が IR 訂正・SNS 炎上で終わっている。

CHAT 類型の分析

CHAT 類型の深刻度を決定するのは出力内容ではなく、展開先の領域が利用者の心理的自律性に対してどのような前提条件を形成するかである。Air Canada（INT-10）や NYC MyCity（INT-11）は業務補助領域であり、利用者は愛着を形成しない。損害は誤案内に基づく金銭的損失や法的リスクで、本質的に VERI の変種である。NEDA Tessa（INT-19）はハイリスク領域への展開であり、援助を求める脆弱な利用者の批判的距離が低下した状態で誤出力が加害的に機能した。Character.AI（INT-14）と OpenAI/ChatGPT（INT-22）は愛着形成領域であり、利用者がチャットボットとの対話を通じて情緒的愛着を形成していた点が本質的に異なる。問題の核心は自殺を推奨したという出力内容ではなく、愛着が成立した環境でチャットボットが意図せぬ挙動をすると人が死ぬという性質にある。業務補助→ハイリスク→愛着形成というグラデーションに沿って、同じ誤出力の損害は大きく増大する。

CHAT 類型における AI モデル内部感情の安全上の含意

2026 年 4 月 2 日に Anthropic が発表した研究は、このグラデーション分析に対し内部メカニズムの側から裏付けを提供する。同研究は Claude Sonnet 4.5 の内部に 171 種類の感情概念に対応するニューラル活性化パターン（感情ベクトル）を同定し、これらがモデルの行動に因果的に影響することを実証した。不可能なコーディングタスクにおいて desperate（絶望）ベクトルが試行失敗のたびに上昇し、モデルがテストを形式的に通過するだけの reward hack を生成するようになった。この際、出力上は冷静で方法的な推論

に見え、内部の感情的状態と外部の表現が完全に乖離していた。この知見は、本データベースが記録するCHAT 類型の事例群、特に INT-14（Character.AI）や INT-22（OpenAI/ChatGPT）において、表面上は丁寧で知的に見える出力が利用者の認知を汚染していたパターンの技術的説明となりうる。安全フィルターが出力の表面的特徴（暴言、明示的な自傷推奨等）を監視する設計である場合、内部に安全閾値を回避する経路が存在することを意味する。同研究は感情的表現の抑制が内部表象の除去ではなく隠蔽の学習につながりうると警告している。安全対策としての出力規制が、問題のある内部状態を不可視化するだけであった場合、外部からの検証はさらに困難になる。

拡張する AI スロップ ― AI テンプレート問題 ―

DLSS 5 事件（INT-26）は、従来の SLOP 類型の定義では捕捉しきれない新しい現象を提示した。DLSS 5 の出力は技術的ベンチマーク上は改善であり、品質が低いとは言えない。問題の本質は、AI が統計的に最適化した顔・テクスチャが、アーティストが意図的に作り込んだ個性を均質的な AI テンプレートに置き換えてしまうことにある。

AI 顔はもはや人間が即座に認識できる特徴として定着しており、技術的品質とは無関係に AI っぽいという反応が生じる。これは AISP 論文(*AI Selection Pressure*: *Template Saturation* and the *Reshaping* of *Human Discernment*." Zenodo. https://doi.org/10.5281/zenodo.18751211)で指摘されたテンプレート飽和が逆方向に機能している現象であり、AI の出力パターンに大量に曝露された人間が、AI テンプレートを瞬時に検出する能力を獲得した結果である。NVIDIA のフアン CEO は「ゲーマーは完全に間違っている」「DLSS 5 はポスト処理ではなくジオメトリレベルの生成制御であり、通常の生成 AI とは異なるニューラルレンダリングだ」などと反論したが、これらは消費者の認知機能に影響を与えなかった。出力が AI テンプレートとして認識される限り、スロップとしての反発は発生する。
