# AI Slop Side Effect Database

## AI スロップ公害事例データベース

低品質な AI 生成物の氾濫によりプラットフォーム・制度が変わり、
正規ユーザー・クリエイター・研究者が間接的損害を受けた事例のデータベース

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

---

## 対象範囲と収録基準

本データベースは AI Incident Database の姉妹版として編纂された。メインデータベースが AI を直接使って損害が出た事例を収録するのに対し、本データベースは低品質な AI 生成物の氾濫によってプラットフォームや制度が変化し、正規のユーザー・顧客・研究者・クリエイターなどが間接的に損害を受けた事例を記録する。

収録基準は次の三つである。①AI スロップの流入が制度や設計変化の原因であること、②プラットフォーム・制度の防衛的変容を経由していること、③具体的な被害者（個人か法人）と実害が確認できること。倫理的に問題がある、今後起きうる、という事象は所見セクションに記述し、事例としては個別に収録しない。

### 類型定義

| コード | 名称 | 定義 |
|---|---|---|
| GATE | ゲートキーピング崩壊 | AI スロップ対策として導入した自動フィルタ・判定ツールが、正規コンテンツ・正規ユーザーを誤排除・誤 BAN した |
| CONT | コンテンツ汚染 | AI スロップの大量流入によりプラットフォームの品質・信頼性が低下し正規ユーザーが不利益を受けた |
| BIAS | 差別的偏向 | バイアスにより特定属性（非ネイティブ・独立研究者等）が系統的に不当排除された |

| | | |
|---|---|---|
| INVIS | 制度的不可視化 | 公式にはオープンを標榜するプラットフォームが不当に検索から排除した |
| COVER | 組織的隠蔽 | AI スロップ対策の誤作動・副作用を指摘された後、組織的に矮小化・隠蔽を図った |

*※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: 学術インフラ領域

プレプリントリポジトリ・学術データベースへの AI スロップ大量投稿が、正規研究者の研究成果発信・可視性・引用に与えた実害。

## PC-1　Zenodo 自動検索除外アルゴリズムと論文「人質」問題

**業界**：学術インフラ / オープンサイエンス

**類型**：GATE + INVIS＋COVER

**時期**：2025 年 11 月〜継続中

**概要:** オープンアクセス論文リポジトリ Zenodo が AI スロップ対策として導入した自動スパム判定アルゴリズムが、手動で認証していないアカウントの論文を検索結果から完全除外する。DOI は有効で OpenAIRE には索引されているにもかかわらず、あらゆる検索で結果が 0 件という人質状態が発生。DOI が存在するため他リポジトリへの投稿は duplicate flag で拒否され、研究者は検索不可視のまま身動きが取れなくなる。さらに AI ツールは論文が存在しないと誤報告し、査読システムが正規 DOI を AI ハルシネーションによる架空文献として処理するという二次被害が発生。Zenodo は「機関に所属すれば解除」などと説明したが、機関が申請している場合においてもサポートは無回答を継続。GitHub issue などで複数の研究者から同症状の報告が集積しており、組織的問題であることが示唆される。被害者の中には AI スロップ問題を専門に研究する研究機関（UTIE Instruments Inc.）が含まれており、AI スロップ対策の副作用が AI スロップ研究者自身を直撃するという深刻な現状を記録する。

Zenodo の公式回答（直接書面）：

> *We have a system at Zenodo to de-rank records and communities which have not yet been manually verified. This system is in place to lessen the impact of spammers. To verify users we expect them to be associated with an institution and using their institutional email and publish quality content.*

***2026 年 3 月 20 日 追記***

Zenodo より公式回答を受領。「Zenodo は学術機関に所属し機関メールを持つユーザーを対象としたリポジトリであり、それ以外からの投稿は検索結果において de-rank されるため、自力での公開や宣伝を推奨」との公式方針が確認された。詳しくは GitHub issue #2604(クローズ済み)を参照。

**損害規模:** 研究成果の検索不可視化、他リポジトリへの投稿不能（duplicate flag）、AI ツールによる論文存在否定、査読プロセスへの支障（複数研究者・継続中）

*関連事案: JP-0（Springer Nature 査読捏造事件）——同一研究機関が AI 査読捏造と AI スロップ対策副作用という二重の制度的障害を受けた記録として関連付ける。*

*出典: Zenodo GitHub issue、Zenodo サポート公式ほか*

## PC-2　arXiv・Preprints.org・SSRN 独立研究者の制度的排除と AI による引用抹消

**業界：** 学術インフラ / プレプリント

**類型：** INVIS + GATE

**時期：** 2024 年〜継続中 ★進行中

**概要:** 主要プレプリントリポジトリがオープンサイエンスを掲げながら実態として一部研究機関以外の研究機関や研究者を排除している問題。直接書面による証拠が複数存在。①arXiv: cs.AI 分野において「査読済み論文でなければ受理しない」「査読誌に掲載されても受理を保証しない」などという通知を送付。プレプリントサーバーとしての本来の定義と完全に矛盾。②Preprints.org: 投稿拒否通知に「scientific merit or quality に基づく判断ではないが」などと主張しながら拒否。③SSRN: 複数研究者の論文が AI 検索ツールに認識されないケースを確認。GPT および Gemini が彼らによって書かれた論文の存在を否定し「架空文献であるから、参考文献から削除すべき」と提案した被害を確認。実際は同論文は DOI で直接アクセス可能であり、Claude Opus および DOI 手動検索では正確に特定。AI システムがプラットフォームの制度的排除を増幅し、存在する論文をハルシネーションとして積極的に消去するという二次被害までも発生したことが確認された。

**損害規模:** 研究成果の流通不能、引用機会の損失、AI による有効な参考文献の削除・引用抹消、研究者や研究機関のキャリア的損害

*本件の核心: AI システムがプラットフォームの制度的バイアスを学習・増幅し、実在する論文を「存在しない」と判定して参考文献から能動的に削除するという行為は、制度的排除の AI 自動化として記録価値が高い。*

*出典: UTIE Instruments Inc.技術資料報告「Designing Beyond Distrust in the Generative AI Era」（2025 年 11 月）*

## PC-3　Stack Overflow モデレーターストライキとコミュニティ崩壊

**業界：** エンジニアリング・コミュニティ / Q&A プラットフォーム

**類型：** GATE + CONT

**時期：** 2023 年 5 月〜継続中

**概要:** 2022 年末に AI 生成回答を全面禁止した Stack Overflow が、2023 年 5 月に AI 検出ツールの使用を禁止するポリシーをモデレーターに通達。AI 検出ツールの誤判定（正確な文章・丁寧な構造・非ネイティブの正式な英語を書く人間の熟練エンジニアが AI と誤認される等）が問題化していたためだったが、これがモデレーターの反発を招き、全モデレーターの 23%以上・Stack Overflow モデレーターの 70%以上がストライキに入った。AI 検出ツールを使うなと AI 回答を排除せよという矛盾した要求の間でコミュニティが機能不全に陥った。2025 年 12 月時点で新規質問数は 2014 年ピーク比 78%減。生成 AI との競合に加え、過剰モデレーション文化がコミュニティを破壊したとされる。

**損害規模:** モデレーター70%以上のストライキ、新規質問数ピーク比 78%減（2025 年 12 月）、エキスパートコミュニティの機能不全

*出典: Stack Overflow 公式発表ほか（2024 年）*

## PC-4　AI 検出ツールの非ネイティブスピーカー系統的誤判定問題

**業界**：学術 / 教育 / 採用

**類型**：BIAS + GATE

**時期**：2023 年〜継続中

**概要:** GPTZero・Turnitin 等の AI 検出ツールが、非ネイティブ英語話者が書いた文章を系統的に AI 生成と誤判定することをスタンフォード大学の研究が実証した。91 本の TOEFL 論文（全て人間が執筆）に対して 7 種の検出ツールが 61.22%を AI 生成と誤判定。うち 19%は 7 ツール全てが誤判定で一致した。ネイティブスピーカーの論文ではほとんど誤判定しなかった。原因は語彙の複雑さが一定、シンプルな文という非ネイティブの特徴が AI 生成のパープレキシティパターンと一致するため。Turnitin は当初「誤判定率 1%未満」と主張していたが、後に 4%に訂正。実際の非ネイティブ文書での誤判定率は 30%以上高い。Vanderbilt 大学では年間 75,000 本の提出論文に対して誤判定率 1%でも 750 名の学生が誤って告発される計算となり、AI 検出ツールの使用を無効化した。複数の大学がツールの使用を停止している。

**損害規模:** 年間数十万人規模の学生・研究者への誤告発、学術処分・奨学金剥奪・精神的被害の報告、非ネイティブスピーカーへの系統的な差別（実証済み）

*特記: スタンフォード研究（Liang et al., 2023, Patterns 誌掲載）により学術的に実証済み。本件は AI スロップ対策ツールが特定の言語的属性を持つ人間を系統的に差別するという重大問題として記録する。*

*出典: Liang et al.「GPT detectors are biased against non-native English writers」Patterns 誌（2023 年）、各大学の公式声明*

## PC-12　Sakana AI「The AI Scientist」による査読リソース浪費・採択枠占有と Nature による正当化

**業界**：学術インフラ / 査読システム / AI 研究

**類型**：CONT

**加害者**：Sakana AI（東京）、University of British Columbia、Vector Institute、University of Oxford。Nature（Springer Nature）は掲載により本行為を学術的に正当化した共同責任者

**時期**：2025 年 3 月（ICLR 2025 投稿）〜2026 年 3 月 26 日（Nature 掲載）

**概要:** Sakana AI（UBC・Vector Institute・Oxford 大学との共同研究）が開発した自動研究 AI システム「The AI Scientist」は、ICLR 2025 ワークショップに対して AI 生成論文 3 本を投稿し、うち 1 本が査読を通過・採択された。Sakana AI は「最初から採択後に撤回する予定だった」などと主張し、実際に採択後に撤回した。AI 生成論文による有限の採択枠の占有、その枠に入れたはずの人間の研究者が排除された可能性、査読リソースの浪費等の加害行為。The AI Scientist のコードは GitHub 上でオープンソース公開されており、1 本あたり 6〜15 ドル・約 3.5 時間で査読通過水準の論文を自動生成できることが実証された。Sakana AI 自身は倫理的配慮などと称して自主撤回を行った。これらの行動は The Supervision Paradox（Naito, 2026）が指摘する V > C_max 動態の直接的な加速要因。PC-6（Clarkesworld）が ChatGPT による大量投稿で投稿受付停止に追い込まれたのと同じことが、学術査読システムにおいて再現される制度的前提が整備された。独立評価論文（Beel et al., 2025）は、The AI Scientist の 42%の実験がコーディングエラーで失敗し、既存概念を新規と誤分類する問題を指摘しており、生成される論文の品質には AI スロップの重大な疑義がある。

**損害規模:**不採択となった人間の研究者への機会損失、査読リソース浪費、オープンソース公開による AI スロップ拡大の間接的支援

**関連事案:** PC-6（Clarkesworld）

*出典: Sakana AI 公式ブログ（2025 年 3 月、2026 年 3 月）、Nature editorial「AI scientists are changing research」（2026 年 3 月）ほか*

# Part 2: クリエイティブ・教育領域

出版・文学・教育領域での AI スロップ大量投稿が、正規クリエイター・作家・学生に与えた実害。

## PC-5 Amazon Kindle AI スロップ汎濫による人間著者市場崩壊と人命リスク事件

**業界**：出版 / 電子書籍

**類型**：CONT

**時期**：2023 年〜継続中

**概要:** 生成 AI により大量の電子書籍が Amazon Kindle に投稿され、ChatGPT で生成した児童書・健康ガイド・クックブック・旅行ガイドが 1 日数百冊レベルで流入した。2023 年には「AI 生成きのこ図鑑」が販売され、毒キノコを食用と誤記した誤情報により人命リスクが指摘され問題化した。Amazon はその後 AI 生成書籍の申告義務・1 日 3 冊の出版上限・手動レビュー強化を導入。しかしこれらの防衛措置により小規模出版社の出版速度が低下し、人間著者の検索順位が AI 書籍の大量流入で埋没し、Kindle ランキングがスパム化するという副作用が発生した。AI スロップが直接の人命リスクを生んだ事例として、プラットフォーム汚染の深刻さを示す。

**損害規模:** 人命リスクを伴う誤情報の流通（AI 生成きのこ図鑑）、人間著者の検索順位埋没、小規模出版社の出版速度低下

*出典: Reuters ほか（2023 年）*

## PC-6 Clarkesworld 誌 AI 短編小説大量投稿による投稿受付停止事件

**業界**：文学 / SF 出版

**類型**：CONT + GATE

**時期**：2023 年 2 月

**概要:** SF 文学誌 Clarkesworld Magazine が、AI 生成短編小説の大量投稿により投稿受付を停止した。編集者が公開した統計では月約 500 本だった投稿数が 1,200 本以上に急増し、その多くが ChatGPT 生成・同一テンプレート構造・プロンプト変形のみという内容だった。編集部は IP ブロック・AI 検出導入・新規投稿停止という防衛措置を取ったが、その結果正規作家の投稿受付が停止し、編集作業量が爆発し、文学誌の通常運営が一時停止に追い込まれた。AI スロップが小規模文学誌のインフラを機能不全に追い込んだ最も記録の明確な事例。

**損害規模:** 正規作家の投稿受付停止、編集作業量の爆発、文学誌の通常運営一時停止

*出典: Wired ほか（2023 年）*

## PC-7 Texas A&M 教授による AI 誤検出告発事件

**業界**：教育 / 大学

**類型**：BIAS + GATE

**時期**：2023 年

**概要:** テキサス A&M 大学の教授が、学生のレポートを ChatGPT に入力して「AI 生成かどうか判定させる」という方法で複数学生を不正行為として告発した。しかし ChatGPT は AI 検出機能を持つツールではないし、判定は完全に不正確。学生の単位が保留され不正行為調査が開始されたが、学生側の抗議を受けて大学は最終的に告発を撤回した。PC-4（Turnitin 等の系統的誤判定）とは異なり、そもそも AI 検出ツールでもないものを AI 検出に使ったという教員の技術リテラシー欠如による誤告発である。メインデータベースの JP-10（三重県虐待死事件）と同じ AI の数値・判定を検証せずに人間が重大な決定を下したという性質だが、本件はよりプリミティブな形で発生。

**損害規模:** 学生の単位保留、不正行為調査開始、精神的被害（大学は最終的に告発を撤回）

*メインデータベース参照: JP-10（三重県虐待死事件）、JP-14（慶應義塾大学レポート不正事件）と合わせて AI リテラシー欠如による教育現場の被害として対照できる。*

*出典: Washington Post ほか（2023 年）*

### PC-8 UC Davis GPTZero 誤判定による学生懲戒委員会付託事件

**業界：** 教育 / 大学

**類型：** GATE + BIAS

**時期：** 2023 年 4 月

**被害者：** William Quarterman（カリフォルニア大学デービス校）

**概要:** カリフォルニア大学デービス校の歴史の試験において、担当教員が AI スロップ対策として導入した GPTZero が、Quarterman の完全な自作答案を AI 生成と誤判定した。この自動判定のみを根拠として彼は当該科目を落第扱いとされ、大学の学術的不正行為懲戒委員会に付託された。後に Google Docs のバージョン履歴（キーストローク記録）を提出することで無実を証明し告発は撤回されたが、プラットフォームが未成熟な検出アルゴリズムに盲目的に依存したことで、無実の学生が AI スロップ生成者の烙印を押されシステムから排除されかけた。PC-7（Texas A&M 事件）と同年の別事件であり、教育現場での AI 検出ツール誤用が複数の独立した事例として頻発していることを示す。

**損害規模:** 成績の不当な取り消し、懲戒委員会への付託による精神的苦痛と名誉毀損（最終的に撤回）

*出典: USA Today ほか（2023）*

---

## Part 3: デジタルプラットフォーム領域

動画・検索・求人プラットフォームへの AI スロップ大量流入が、正規クリエイター・事業者・フリーランサーに与えた実害。

### PC-9 YouTube AI 児童向けスロップ汚染問題

**業界:** 動画プラットフォーム / 教育

**類型:** CONT

**時期:** 2023 年〜継続中 ★進行中

**概要:** YouTube および YouTube Kids に AI 生成の児童向け動画が大量流入した。New York Times（2025 年 12 月）の調査報道によれば、AI で生成された教育を装う動画が YouTube の推薦アルゴリズムにより子どもに積極的に配信されていた。内容は教育的価値が皆無であるだけでなく、既存の絵本を無断で AI 動画化して単語の発音を誤って教える事例や、人気キャラクターを暴力的な場面に登場させる不適切動画も含まれていた。ミシガン大学の発達行動小児科医 Jenny Radetsky は認知発達への悪影響を指摘。Pew Research Center の調査では 2 歳未満の子どもの約 60%が YouTube を視聴しており、約 3 分の 1 が毎日視聴している。YouTube は AI 生成のリアルなコンテンツには開示を要求しているが、児童向けのカートゥーン調 AI 動画は開示義務の対象外であり、最も子どもに到達しやすいコンテンツが制度的に捕捉されないという矛盾が生じている。NYT 報道を受けて 5 つのチャンネルが YouTube パートナープログラムから除外されたが、根本的な対策には至っていない。欧州委員会もデジタルサービス法に基づく児童保護の観点から YouTube を調査中。

**損害規模:** 児童への認知発達リスク、正規教育コンテンツクリエイターの埋没、絵本出版社の著作物無断 AI 動画化

*出典: New York Times（2025 年 12 月）ほか*

## PC-10 フリーランスプラットフォーム AI 代替によるフリーランサー収入崩壊

**業界:** フリーランス / ギグエコノミー

**類型:** CONT

**時期:** 2023 年〜継続中 ★進行中

**概要:** この事例は単なる技術的代替と AI スロップ公害の両面があるため、両者の切り分けは今後のより一般化できる実証的研究が待たれる。Upwork・Fiverr などのフリーランスプラットフォームにおいて、AI ツールの普及によりフリーランサーの契約数・収入が系統的に減少している。Brookings 研究所が引用した Organization Science 誌掲載の研究（Hui et al., 2024）は、Upwork のデータを分析し、生成 AI への露出度が高い職種のフリーランサーが契約数 2%減・収入 5%減を経験したことを実証した。特に**高品質・高価格帯のサービスを提供していた熟練フリーランサーへの影響が顕著**だった。Sensor Tower のデータによれば、2024 年上半期に Fiverr のダウンロード数は前年比 18%減、Upwork は 22%減。Upwork のフリーランサー向けアプリの月間アクティブユーザーは 5 四半期で 32%減少した。Ramp 社の経済研究（2025 年）では、フリーランスマーケットプレイスへの企業支出シェアが 2021 年 Q4 の 0.66%から 2025 年 Q3 の 0.14%に急落する一方、AI モデルプロバイダーへの支出は同期間にゼロから 2.85%に急増。フリーランス支出 1 ドルの削減に対し AI 支出はわずか 0.03 ドルという置換比率が示された。Fiverr 社は 2024 年にアクティブ購入者数が前年比 10%減少した。本件はプラットフォーム自体が AI スロップで汚染されたというよりも、AI の存在がプラットフォーム利用の経済的前提を崩壊させ、正規フリーランサーの生存基盤を掘り崩した事例であり、AI スロップ公害のアナロジーが成立する。

**損害規模:** Upwork フリーランサー向けアプリ MAU 32%減（5 四半期）、Fiverr アクティブ購入者 10%減（2024 年）、AI 露出度が高い職種で契約数 2%減・収入 5%減（学術実証済み）

*出典: Brookings 研究所（2025 年 7 月）、Hui et al. Organization Science 誌（2024 年）、Sensor Tower、Ramp 経済研究（2025 年）*

## PC-11 Google 検索 AI 汚染・正規サイトトラフィック崩壊

**業界:** 検索プラットフォーム / デジタルメディア

**類型:** CONT + GATE

**時期:** 2024 年〜継続中 ★進行中

**概要:** AI スロップの SEO スパムとしての大量流入および Google の AI Overview（AI 生成要約）の導入により、正規ウェブサイトのオーガニックトラフィックが大規模に崩壊した。ライプツィヒ大学等の研究チーム（Bevendorff et al.）による 1 年間の実証研究で、Google 検索結果における低品質 SEO スパムの増加と、検索アルゴリズムによるスパム対策の一時的効果しか持たないことが確認された。Google は 2024 年 3 月のコアアップデートで低品質コンテンツを 40%削減すると宣言し、AI コンテンツファームを含む数千のサイトが 60-80%のトラフィック減を経験した。しかし、Google の防衛策は同時に AI スロップとは無関係な正規サイトをも巻き込んだ。AdExchanger（2026 年 1 月）は Business Insider がオーガニック検索トラフィック 55%減・従業員 21%削減、音楽ブログ Stereogum が広告収入 70%減、旅行ブログ The Planet D がトラフィック半減と報じた。小規模ホームインプルーブメントブログ Charleston Crafted は 2024 年 3-5 月にトラフィック 70%減・広告収入 65%減。HuffPost もデスクトップ・モバイルの検索リファラルが半減した。2024-2025 年の検索全体で 60%以上がゼロクリック検索（ユーザーがリンクをクリックせず離脱）となっており、AI スロップの流入→Google の防衛的アルゴリズム変更→正規サイトの巻き添え損害という本データベースの中核的構造がインターネット全体規模で発

生している。

**損害規模:** Business Insider オーガニック検索トラフィック 55%減・従業員 21%削減、Stereogum 広告収入 70%減、Charleston Crafted トラフィック 70%減・広告収入 65%減、HuffPost 検索リファラル半減、複数の小規模パブリッシャーの閉鎖

*出典: AdExchanger（2026 年 1 月）、Bevendorff et al. ライプツィヒ大学研究、Bloomberg、Seer Interactive 分析、Grow and Convert 分析*

### PC-13 Reddit API ポリシー変更・モデレーターストライキ事件

**業界:** ソーシャルプラットフォーム / コミュニティ

**類型:** GATE + CONT

**時期:** 2023 年 6 月〜継続中

**概要:** Reddit が AI 企業による大量スクレイピング対抗策としてサードパーティ API 利用を有料化したことで、主要サードパーティクライアントが相次いで閉鎖。これに反発した 8,000 以上のサブレディットが 2023 年 6 月にブラックアウト（一時閉鎖または非公開化）を実施し、r/funny など大規模コミュニティを含む一部は無期限閉鎖を宣言した。障害を持つユーザーが代替クライアントのアクセシビリティ機能に依存していたケースでは実害が直接的であった。PC‑3（Stack Overflow）との同一性は明確であり、AI スクレイピング流入→プラットフォームの防衛的ポリシー変更→正規ユーザー・モデレーター・サードパーティ開発者が巻き添え被害を受けるという本データベースのパターンが、インターネット最大規模のコミュニティプラットフォームで発生した事例として記録する。PC‑3 がコンテンツ判定の矛盾で崩壊したのに対し、本件はアクセス経路の物理的遮断という防衛手段が正規ユーザーのインフラを直撃した点で補完的な事例類型となる。

**損害規模:** 8,000 以上のサブレディットのブラックアウト、主要サードパーティクライアントの閉鎖、障害者ユーザーのアクセシビリティ喪失、モデレーターの大規模離脱

*出典: The Verge・Wired 等国際主要メディア報道（2023 年 6 月）*

## 所見

### AI スロップ公害とは

本データベースが記録する事例群に共通する点は、環境公害のアナロジーで理解できる。工場が有害物質を垂れ流すと川が汚染され、何も悪いことをしていない下流の住民が被害を受ける。AI スロップを大量生成する者が垂れ流すとプラットフォームが防衛的になり、何も悪いことをしていない正規ユーザーが巻き込まれる。PC-4 の非ネイティブスピーカーは自分で AI を一切使っていない。人間が書いた文章を AI 検出ツールが誤判定し、学業や研究キャリアが破壊される。PC-1・PC-2 の研究者は AI スロップを一切投稿していない。他人のスロップに対するプラットフォームの防衛反応により、検索から消されるか、AI に「この論文は存在しない」と抹消される。メインデータベースが記録する事例でも、被害者は AI とかかわらなければよいという選択肢はそもそも持っていないことも多くある。メインデータベース INT-23 の高齢患者は保険会社が導入した AI に給付を打ち切られ、INT-1 の消費者は知らないうちに AI に価格を 23%上乗せされ、JP-9 の 4 歳児は AI の存在すら知らないまま死亡した。AI を使っていない人間が AI によって損害を受け、しかも自分が AI に損害を受けたことすら認識できない。これが我々が、AI のインシデントについて公害と呼ぶ根拠である。

## 今後の拡大が予測される領域

プラットフォームが AI スロップ対策として導入する自動フィルタは、必然的に保守的な閾値設定になる。スロップを見逃すよりも正規コンテンツを巻き込む方が安全という組織的判断が働くためである。PC-3（Stack Overflow）のモデレーターストライキはこの矛盾が爆発した典型であり、PC-4（AI 検出ツールの系統的誤判定）・PC-7（Texas A&M 誤告発）と合わせて「AI 検出ツール 3 兄弟」として参照できる。現時点では具体的被害者の確認が不十分なため本データベースには収録しないが、以下の領域で同構造の問題が拡大することが予測される。①音楽・映像プラットフォームでの誤 BAN（Bandcamp 等が AI 禁止ポリシー導入済み、誤判定発生は時間の問題と予測）、②文学・絵画コンクールでの AI 検出誤判定による失格(既に SNS では多数報告あり)、③就職応募・グラント審査での AI 生成書類大量投稿による審査プロセス形骸化(既に SNS では多数報告あり)、④特許出願への AI ツール利用拡大による審査インフラ過負荷（2025 年 12 月の日本の特許出願件数が 8 万 2188 件と前月比約 2.69 倍に急増。AI 特許支援ツールによる出願コスト削減が背景とされ、SNS では「特許庁への DDoS 攻撃」と表現する声も出た。国立情報学研究所の佐藤一郎教授は一部の企業や特許事務所が AI ツールを使って大量出願した可能性を指摘。AI 特許支援ツールベンダーの CEO は「従来の人的リソースのみでは現実的に処理が困難な規模」と認めた。特許庁の審査官の認知処理能力 C_max を超える出力速度 V が発生した事例として、Supervision Paradox 論文が予測する V > C_max 動態の制度インフラ領域での実証例となりうる。ただし現時点では特許庁側の防衛的変容および具体的被害者の特定が不十分なため、所見にとどめる。出典: 日経クロステック 永田雄大（2026 年 3 月）「12 月に急増した特許出願件数の謎、コスト削減につながる AI 利用拡大が背景に」「特許出願数が異例の水準に、25 年 12 月は前年同月比 170%増 ちらつく AI の影」）。これらは具体的被害者が確認され、より強いエビデンスと損害が見つかり次第、本データベースに収録する。

⑤学会誌への AI 論文判定ソフト導入による正規論文の誤排除懸念（2026 年 3 月、国立情報学研究所の越前功教授らが開発した AI 論文判定ソフトを日本物理学会が学会誌の査読に導入。NHK が報道。精度 90%と報じられたが、投稿量が通常水準である検証環境での数値であり、AI 生成投稿の大量流入により投稿量が 2 倍以上に膨張した実運用環境での偽陽性率は未検証と推定される。また非ネイティブ英語話者の論文を AI 生成と誤判定する傾向は複数の AI 検出ツールで報告されており（PC-4 参照）、日本の学会では投稿者の大半が非ネイティブであるため環境固有のリスクが高い。精度の数値的アピールが投稿量一定という暗黙の前提に依存している点は、本論文 Section 4.3「Error Rate Improvement Does Not Guarantee Safety」が示したものと同一であり、実際に処分を含む運用がなされないか懸念。出典: NHK ニュース 2026 年 3 月 25 日）

## メインデータベースとの関連

AI Incident Database（メインデータベース）の JP-0（Springer Nature 査読捏造事件）の被害者である UTIE Instruments Inc.は、本データベース PC-1（Zenodo）・PC-2（arXiv 等）の被害者でもある。**AI スロップ問題を研究する機関が、AI スロップと AI スロップ対策の副作用という両方の損害を短期間で受けたという事実**は、本データベースが記録する問題の深刻さを象徴している。

## 被害者は声を上げられない

本データベースに収録された PC 事例（Zenodo・Stack Overflow・Clarkesworld 等）が表面化した理

由は、被害者の中にすでに実績と証拠を持つ告発者がおり、不当な排除に対して声を上げる手段を持っていたからである。しかし採用選考・助成金審査・フリーランスプラットフォーム・金融与信といった領域では、AI スロップ対策の誤作動による被害がはるかに広範に、かつ完全に不可視のまま進行している可能性が否定できない。AI 検出アルゴリズムの偽陽性によって機械的に足切りされたとしても、通知されるのは「人間による慎重な検討の結果、誠に残念ながら」という定型文のみである。被害者は「自分の実力が足りなかった」「運が悪かった」と思い込み、一人で絶望する。異議申し立ての窓口は存在せず、被害者は自分が差別されたことすら知らない。被害者が被害を認識できないという点で、これは本データベースに収録された事例より深刻な問題といえるだろう。

## COVER（隠ぺい）は経済的合理性として発生する

なぜ企業・プラットフォームは度々AI スロップ公害インシデントを隠蔽するのか。それはモラルの問題ではなく、いくつかの合理的帰結である。まず、総合的な判断という法的シールドの存在である。採用・与信・モデレーションにおいて、企業には判定プロセスの詳細を外部に開示する法的義務がない。「社内規定に基づく総合的な判断」という定型文を使えば、それが AI 検出ツールの誤作動であろうと非ネイティブへの差別であろうと、すべてを合法的に隠ぺいできてしまう。そして、認めた瞬間に発生するシステマティックな差別の立証リスクである。もし企業が AI 検出ツールの誤判定と認めた場合、それは同じアルゴリズムで弾かれた過去の全員に対するシステムレベルの差別を公式に自白したことになる。これは、集団訴訟・コンプライアンス違反に直結するため、法務部は「絶対にシステムエラーとは認めるな」と被害を個人の不満へ矮小化する組織的隠ぺいモードを強制する。AI 判定を人間が覆すことの経済的非合理性もある。コスト削減のために導入した AI 判定を人間が精査し直すことは導入の目的そのものを否定する。企業にとって無実の人間を数人誤って捨てるコストより、判定プロセスに人間を介入させるコストの方が圧倒的に高い。異議を唱えられた時だけ後付けで、人間がちゃんと確認した(ヒューマンインザループ)という建前を引っ張り出す。本データベースのメインデータベースに収録された Springer Nature らの隠蔽パターン（COVER 類型）と本質的な理由は同じであり、AI の出力が表面的には正しそうに見えるため管理職が「適当に取り繕えば逃げ切れる」などと判断し、組織的隠蔽が最もコストパフォーマンスの良い選択肢として採用される。

## 差別の連鎖

これが従来のスパム問題との最も本質的な違いである。昔のスパムフィルターが誤作動しても、そのメールが相手に届かないだけで終わった。しかし現在は、一つのプラットフォームが正規ユーザーを検索から除外すると、そこを参照・学習している他の AI システム（検索 AI・対話 AI・引用チェックシステム等）が「その人物・その企業・その論文は存在しない」と学習する。PC-2（arXiv・SSRN 問題）で記録した通り、AI システムが実在する論文を「架空文献である」と判定して参考文献リストから自動削除するという事態がすでに発生している。一つのプラットフォームの雑な防衛アルゴリズムの誤作動が、そこを参照する世界中の AI システム全体に連鎖・増幅し、実在する人間の業績と実存をデジタルナレッジから自動的に抹消していく。被害は一つのプラットフォームで完結せず、デジタル世界全体に波及する。これが、AI スロップ公害が単なる技術的ミスを超えた社会的問題になりうる理由である。

## フロー設計アプローチ

本データベースが記録する AI スロップ公害への現行対応は、いずれも人間による監督強化アプローチの失敗パターンに収束している。スパムが増えた→チェックを増やす→人間の処理限界を超える→形骸化→雑な制度の実装と運用で正規ユーザーを巻き込む、という性質である。PC-1（Zenodo）・PC-3（Stack Overflow）・PC-4〜PC-8（各 AI 検出ツール誤判定事例）はすべてこのパターンの帰結である。「The Supervision Paradox: AI Capability Growth Necessitates Usage Contraction in High-Loss Domains」（Naito, 2026）では、この失敗の根拠を三つの同時成立する制約から導出している。責任は人間に帰属する（R=1）、人間の認知処理能力には生物学的上限がある（C_max）、経済的圧力により AI の出力速度は非線形に拡大する（V→∞）。この三制約が同時に成立する限り、監督強化は持続可能な均衡を形成しない。出力速度 V を人間の処理限界以下に制限すること（V_eff = min(V, C_max)）が期待損失を有界に保つ唯一の政策変数として導出される。このフロー設計アプローチをプレプリントリポジトリや投稿プラットフォームのスパム問題に適用すると、以下の原則が導かれる。まず、コンテンツの質・真正性を人力や AI で検知しようとする試みは全て無駄である。それはさらなる偽陽性を生み、あるいはモデレーションの速度が物理的に投稿速度に追いつけない。解決策は物理的な投稿速度の制限のみである。次に、KYC（パスポート・納税証明・政府発行 ID など、機関メールに限定しない一般的な身元確認）済みの一人間につき、一定期間あたり N 件の投稿権枠を付与する。N の値は分野・プラットフォームの適正スループットに依存するが、真正な研究者の通常の生産ペースを下回らない水準に設定する。最後に、この物理的制約下にあるコンテンツはコンテンツ内容による AI 自動判定やモデレーションの対象外（聖域）とする。この設計の下ではスロップ生成者は経済的インセンティブを失い、正規ユーザーは聖域を得る。

なお、その論文「The Supervision Paradox: AI Capability Growth Necessitates Usage Contraction in High-Loss Domains」（Naito, 2026）自体が、arxiv への投稿において人間の AI 分野の専門家(endorser)からの推薦状を取得・法人研究機関の実用的にも理論的にも有効な研究であるにもかかわらず、専門知識のない arxiv のモデレーターにより on-hold のまま音沙汰なく長く放置されるという事態に直面しており（PC-2 参照）、arxiv のシステムが最も必要とする解決策を提示した研究機関と著者(当機関の主任研究員)を排除するという深刻な問題を現在進行形で記録している。

このフロー設計は既存の信頼階層を破壊しない。むしろ両立させることで制度移行コストを最小化し、既得権益層の制度的抵抗を排除できる。既存の大手研究所・大手企業所属の研究者は現行通り制限なしとする。有名大学・大手出版社・実績のある研究者にとって自分たちの特権的地位は維持されるため制度的抵抗が最小化される。むしろ一般ユーザーとの差別化が明確になるという点で、現在の体制が続くことになり、歓迎する可能性が高い。そして、KYC 済みの全ての個人・法人に対して量的制限付き投稿枠を保証することで、この枠内のコンテンツはコンテンツ内容による AI 自動判定の対象外（聖域）とする。スロップ生成者は限界費用ゼロで無限投稿できるという経済的前提が崩れるため ROI がゼロになり、自然淘汰される。これは、制度設計上の利害対立を最小化しながら AI スロップ公害の根本原因に対処できる現状で唯一の解である。

## なぜ正しいアプローチが採用されないのか

現在の arXiv・Zenodo 等のプラットフォームは、いわば蛇口を全開にして汚水を垂れ流しながら、下流でフィルターが目詰まりしたと騒いでいる状態である。なぜ蛇口に KYC という物理的流量制限をつけないのか？この問いに対して Naito,(2026)は既に 3 つの理由を議論している。第一に、経済的インセンティブの非対称性である。蛇口を開けること（投稿の自由化・ユーザー数増加）はプラットフ

ォームの成長として短期的に評価される指標である。KYC を導入すれば摩擦が増えユーザー数が減りインプレッションが落ちる可能性がある。被害者（正規研究者）は声が小さく、受益者（スロップ業者）は目に見えない。蛇口を締めることの短期的コストは明確だが、締めないことのコストは不可視である。そして、スロップ汚染の被害は累積するが、臨界点を超えるまで表面化しない（E_detected ≪ E_actual）。現場の担当者は「まだ大丈夫」と認識し続ける。PC-6（Clarkesworld）のように突然崩壊するまで問題は静かに蓄積しているだけである。フィルターが目詰まりして、初めて大変なことになったと騒ぐのは臨界点を超えた後の反応的対応（論文での Threshold Shock、Phase III: Reactive などを参照）である。根本的な問題は、蛇口を締める設計をした人間が責任を取る必要がないという問題である。プラットフォームを設計した人間はオープンサイエンス・言論の自由という理念を掲げており、その理念を守っている限り批判されにくい。しかしその理念の結果としてスロップが流入し正規ユーザーが排除されても、その因果関係を自分たちの設計ミスとして認めると、PC-1・PC-2 で整理した通りクラスアクションのリスクがある。だからスパム業者が悪い、AI が悪いという矮小化や隠ぺい工作（COVER）をするのが合理的選択になる。

## arXiv 上の企業研究者の実態：大企業は別ルートを持ち、中小は排除される

Google DeepMind・Microsoft Research・Meta AI・OpenAI 等の大企業研究者は、arXiv と自社ブログ・学術会議を並行して使うマルチチャンネル戦略を取っている。arXiv へのプレプリント投稿と同時に自社ドメインで今週の研究ハイライトとして公開し、学術会議採択後は arXiv リンクと自社ブログ記事を両方展開する。彼らにとって arXiv は必要に応じて使うチャンネルの一つであり、on hold になっても自社ドメイン・Google Scholar・学術会議という代替経路が無数にあるため実質的なダメージがない。ファーウェイをはじめとする中国系大企業も arXiv への大量投稿を行っており、機関メールの特権で自動通過できる。arXiv が「AI スロップで一番打撃を受けているのは CS カテゴリ」と認め、レビュー論文・ポジションペーパーを原則禁止にした直接の原因の一つがこの大量投稿である。皮肉なことに大企業が AI スロップを垂れ流し、機関所属のない独立研究者・小規模法人がその防衛措置の巻き添えを食らうという現象が発生している。結果として現在の arXiv は以下の四層構造になっている。第一層の GAFA・DeepMind 等は自動通過・大量投稿が可能で自社ドメインでも補完できる。第二層のファーウェイ等中国大企業は機関メールで通過し**スロップ混在のまま大量投稿**している。第三層の中規模企業研究者はグレーゾーンで on hold が多発する。そして第四層の小規模法人や個人研究者は推薦状を保有しているにもかかわらず on hold による実質的に排除がされている。これはオープンアクセスの看板を掲げながら、実態としては大企業の研究インフラとして機能しているシステムの姿である。

## KYC こそが真に民主的なオープンインフラの条件

現行プラットフォームが採用している所属機関名＝信頼という認証モデルは、表面的にはオープンに見えて実態は大手機関所属者を優遇する貴族趣味的な設計である。MIT のメールアドレスを持つ人間は無検証で信頼され、パスポートしか持たない個人研究者は疑われる。これはオープンサイエンスの理念と真逆の帰結である。公的な KYC（納税証明・パスポート・政府発行 ID 等）が問うのは、**どこの組織に属しているかではなく、責任を負える一人の実在する人間か**だけである。この問いこそが、AI スロップ時代における真に民主的でオープンなインフラの唯一の正当な条件である。AI スロップが問題なのは、それが大量かつ無責任に生成されるからである。逆に言えば、責任を負える実在する

人間が、自分の投稿に責任を持つという条件さえ満たせば、その人間が大学に所属していようとなかろうと、内容の真正性とは無関係に投稿の権利を持つべきである。大手研究機関所属を問うことは、この問いへの回答にならない。AI はパスポートを持てないし、納税義務を負えない。しかし実在する人間はそれができる。この非対称性こそが、KYC が AI スロップと人間の知性を分ける唯一の正当な検問所である根拠であると我々は主張する。**オープンサイエンスを標榜しながら利益相反関係であることを事実上の条件とする現行プラットフォームは、KYC という民主主義の条件を拒否することで、既得権益層の特権を開放という言葉で偽装している。**フロー設計アプローチにおける KYC は、排除の道具ではなく包摂の道具である。世界中のいかなる研究者も、人間としての実在さえ証明できれば等しく聖域枠を持つことができる。これは現行システムが達成できていない、そして達成する気もない真のオープンサイエンスの姿である。

## AI スロップ公害からのプラットフォーム防衛

フロー設計アプローチの必要性は学術インフラに限定されない。Part 2・3 の事例群は、同じ失敗がクリエイティブ・教育領域およびインターネットの基幹インフラにおいても発生していることを示している。PC-5（Amazon Kindle）では、AI 生成きのこ図鑑が毒キノコを食用と誤記し人命リスクが発生してから、Amazon は 1 日 3 冊の出版上限と手動レビュー強化を導入した。PC-9（YouTube）では、AI 生成児童動画が推薦アルゴリズムで 2 歳児に積極配信されていることが NYT に報じられてから、5 チャンネルをパートナープログラムから除外した。いずれも Supervision Paradox 論文の Phase III（反応的対応）そのものであり、蓄積期には何も検知できず対応が間に合っていない。しかも YouTube の開示義務はなぜかリアルに見えるコンテンツのみを対象としていたため、リアルかどうかを判断する能力を持たない 2 歳児向けのカートゥーン調 AI スロップはもとから捕捉されない設計であった。PC-11（Google 検索汚染）は最も大規模な実証である。Google は AI スロップの SEO スパム流入に対して 2024 年 3 月のコアアップデートで低品質コンテンツを 40%削減すると宣言した。結果として AI コンテンツファームは打撃を受けたが、同時に Business Insider がオーガニック検索トラフィック 55%減で従業員 21%を削減、音楽ブログ Stereogum が広告収入 70%減、小規模ブログ Charleston Crafted がトラフィック 70%減・広告収入 65%減という正規サイトへの巻き添え被害が発生した。これはフロー設計アプローチが主張する「コンテンツの質・真正性を判定しようとする試みは全て無駄、さらなる偽陽性を生むか、モデレーションの速度が投稿速度に追いつけない」の実証例である。Google という世界最高水準の研究力と資金を持つ企業が、世界最大規模のリソースを投じて内容判定アプローチを採ってなお、正規サイトの大量誤排除を防げなかった。これら 3 つの領域が合流する結論は一つである。プラットフォームの規模も技術力も関係なく、コンテンツの中身を判定して良質と粗悪を選別するアプローチは破綻する。物理的な流量制限だけが、偽陽性を生まずに AI スロップを排除できる。

**2026 年 4 月 13 日追記:** 我々のフロー設計アプローチの妥当性が、予想外の形で実証された。2026 年 4 月 7 日、Anthropic は汎用 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・推論能力の向上だけであらゆる主要 OS・ブラウザのゼロデイ脆弱性を自律的に発見・エクスプロイトする能力を獲得したと発表した。Anthropic は同モデルを一般公開せず、約 40 社に限定アクセスを提供する Project Glasswing を開始。これはまさに Supervision Paradox 論文（Naito, 2026）が結論づけた「出力速度 V が人間の監督能力 C_max を超えた場合、V_eff = min(V, C_max)として出力速度自体を制限するしかない」というフロ

ー制限の現実世界での実装そのものである。コンテンツの中身を判定して良質と粗悪を選別するアプローチはやがて破綻するという当社の一貫した主張は、サイバーセキュリティという物理的障壁が少なくかつ高損失の領域において最初に成立した。Anthropic はモデルの出力の中身を判定して安全な使い方だけを許可するアプローチではなく、アクセスできる人類の数を物理的に制限するフロー制限を採用した。Anthropic のフロンティアレッドチーム責任者は 6〜18 か月以内に他社から同等能力のモデルが登場すると予測しており、フロー制限による時間稼ぎは本質的に一時的措置に過ぎない。しかし重要なのは、現在世界最高の AI モデル開発企業が、自社の最新モデルの安全性担保として採用した防御策がコンテンツ判定ではなくフロー制限であったという事実である。彼らは自分たちが作ったモデルの出力を既に人間のチェックの強化では安全に判定できないことを知っている。これは本データベースが記録してきた PC-1（Zenodo）・PC-3（Stack Overflow）・PC-4〜PC-8（AI 検出ツール誤判定）・PC-11（Google 検索）の全てに共通するコンテンツ判定型アプローチの失敗のメタ的な証明であり、フロー設計アプローチが学術インフラからサイバーセキュリティまで領域横断的に時間稼ぎとしては妥当であることを示唆している。（メインデータベース INT-30 参照）