AI による被害のメカニズム

AI インシデントから得た教訓

エグゼクティブサマリー

生成的人工知能、特に強力な生成モデルの最近の進歩に伴い、民間・公共セクターの関係者は、AI を日常生活により顕著に組み込むことの利点を称賛している。頻繁に言及される利点には、生産性、効率性、パーソナライゼーションの向上が含まれる。しかし、AI が引き起こす害については、まだ十分に理解されていない。 AI の普及と利用拡大に伴い、近年 AI による被害インシデントが急増しており、現行の被害防止策が不十分である可能性を示唆している。本報告書は、この原因が AI リスクが実際にどのように顕在化するかの理解不足にあると論じる。AI インシデントデータベースの事例報告を活用し、AI 展開が被害をもたらすメカニズムを分析し、このプロセスを説明する 6 つの主要なメカニズムを特定した（表 1）。

表 1：6 つの AI 危害メカニズム

意図的な危害	意図しない危害
● 設計による危害 ● AI の悪用 ● AI システムへの攻撃	● AI の失敗 ● 人間の監視の失敗 ● 統合的な危害

これらのメカニズムに関連する AI インシデントを検証すると、将来の AI ガバナンスアプローチに活かすべき重要な示唆がいくつか得られる。

1. **危害防止への画一的なアプローチは不十分である。**本報告書は、AI による危害に至る多様な経路と、関与する幅広い主体を示している。効果的な緩和には、社会技術的アプローチを含む同様に多様な対応戦略が必要である。モデルベースのアプローチのみを採用すると、特に統合上の危害や人的監視の失敗を見落とす恐れがある。

2. **これまでのところ、危害リスクはモデルの能力と弱い相関しか示していない。**本報告書は、単一目的の AI システムが関与する多くの危害事例を明らかにしている。にもかかわらず、多くの政策アプローチでは、計算能力で代用される広範なモデル能力を、危害発生傾向の予測因子として用いている。これは、能力の低い AI システムの無責任な設計・開発・展開に伴う重大なリスクを緩和できていない。

3. **AI インシデントを追跡することは、現実の AI リスクに関する貴重な知見を提供し、対応能力の構築に役立つ。**技術革新、新たなユースケースの実験、斬新な攻撃戦略は、将来新たな AI 被害インシデントを引き起こすだろう。こうした進展に対応するには、迅速な適応と機敏な対応が求められる。包括的な AI インシデント報告は、 加速されたペースでの学習と適応を可能にし、改善された緩和戦略と新たな AI リスクの出現時の特定を可能にする。インシデント報告は、AI リスクに対処するための重要な政策ツールとして認識されなければならない。

目次

エグゼクティブサマリー **1**

序論 **3**

方法論 **4**

制限事項 5

AI による危害のメカニズム **6**

意図的な危害 6

意図しない危害 9

考察 **14**

結論 **15**

附属書 **16**

謝辞 **18**

脚注 **19**

序論

AI の普及と展開が進むにつれ、AI システムは有害な事象に関与するケースが増えている。 2025 年の初めだけで、AI インシデントデータベース（AIID）に 279 件の新規インシデントが追加された。AIID は非営利団体による取り組みで、AI 展開による現実の被害を追跡することを目的としている。[1] 2020 年 11 月の立ち上げ以来、このデータベースはアルゴリズムシステムや AI に関連する被害やニアミス事例を 1,200 件以上収集・索引化している。[*]

明らかに、こうした AI による被害を防ぐためにはさらなる取り組みが必要である。 欧州連合の AI 法のような規制、行政指導である米行政管理予算局覚書 M-25-21、Anthropic 社の責任あるスケーリング方針のような企業枠組みを問わず、ほとんどの AI ガバナンス介入が追求する根本的な目標は、被害の事前防止である。[2] しかし、被害を効果的に防止するには、理論上ではなく実践において、AI の使用がどのように有害な結果をもたらすのかをより深く理解する必要がある。

AI インシデントデータは、AI システムが実際にどのような害を引き起こすかを理解する上で貴重な知見を提供する。AIID は数百件の実世界の AI インシデント報告を収集・索引付け・アーカイブすることで、AI システムが関与した無数の危害だけでなく、それらの危害がどのように発生したかを記述する貴重なデータ群を構築した。 CSET は以前、このデータを活用して、インシデントデータ分析のための基本的な定義と分類体系を提供する分析枠組みを作成した。[3] この枠組みはその後、データベースから 200 件以上のインシデントを、インシデントタイプ、危害カテゴリー、その他の次元で注釈付けし分類するために使用された。[4]

この過去の分析作業が本報告書の基盤となっている。本報告書は AI による被害に関わる多様な要因を記述するものである。データベースから検証した 200 件以上の実世界の被害事例を活用し、意図的な被害と意図しない被害に分類できる 6 つの主要な「被害メカニズム」を特定した（表 2）。

---

[*]AIID 内の各インシデントは 1 つ以上の危害事例に対応するため、データベースに記録された個別の危害事象の総数はインシデント ID 数よりも多い。一部のインシデント ID は単一の記録された危害事例に対応するが、他は関連するインシデントを単一の物語に集約したメディア構築の記述を捕捉している。

表 2：AI による 6 つの危害メカニズム

意図的な危害	意図しない危害
● **設計による危害：**有害な目的で設計・開発された AI システムによる危害 ● **AI の悪用：**開発者の意図に反する危害を及ぼすための AI システムの使用 ● **AI システムへの攻撃：**サイバー攻撃によって引き起こされた AI の行動または（不）作為による危害	● **AI の失敗：**AI のエラー、誤動作、またはバイアスによって引き起こされる危害 ● **人間の監視の失敗：**人間と機械のチームが機能しなかった結果生じる危害 ● **統合的な危害：**特定の文脈での展開が意図せぬ結果として引き起こす危害

これら 6 つのメカニズムは、AIID で特定された様々な被害経路を包括的に説明する。したがって、AI リスクが実際にどのように顕在化するかをより深く理解する助けとなり、緩和策の指針となる。

これらの危害メカニズムは、欧州連合の AI 法のような規制への準拠を目指す企業にとって、即時の政策関連性を持つ可能性がある。EU は最近、汎用 AI のための行動規範を発表した。この自主的な準拠ツールは、汎用 AI モデルプロバイダが同法の要件を遵守するのを支援するために開発され、包括的なリスクマネジメントプロセスを定めている。 これに基づき、開発者はリスクモデリングに取り組む必要がある。これは「モデルに起因するシステミックリスクが顕在化する経路を特定することを目的とした構造化されたプロセス」と定義される [5]。実世界のインシデントに基づく実証的証拠の上に構築されたこの危害メカニズムの枠組みは、この取り組みの有用な出発点となり得る。

重要なのは、これらの多様なメカニズムに対処するには、同様に幅広い緩和戦略が必要であるということだ。セキュリティ対策は悪用や攻撃のリスクを軽減できるが、統合的な危害には効果がない。性能標準やテスト手順は AI の失敗を減らせるが、人間の監視の限界を緩和することはできない。AI による危害を効果的に防ぐには、多様なツールボックスが必要である。

以下の節では、AIID が収集した被害インシデントを提示し、6 つの被害メカニズムを説明するとともに、AI が被害をもたらす多様な方法を理解を深める。本報告書は緩和技術の包括的概観を提供するものではないが、可能な範囲で特定のメカニズムに対抗する対策を強調し、特定の課題に対する効果的な緩和策を見つけるためにさらなる研究が必要な領域を記述する。

方法論

本報告書の分析は二段階で実施された。第一段階では、AI インシデントの詳細な研究に基づき、六つの危害メカニズムを特定した。先行研究において CSET は、AI 危害の研究のための標準分析枠組みを開発している。[6] この枠組みの開発には、インシデントのレビューと枠組みの適応を繰り返す反復プロセスが含まれ、現実世界の AI 危害を特定するために必要な主要要素、それらの基本的な関係構造、および定義が明らかになった。 この枠組みの な中核概念は「危害の連鎖」である。これは AI の展開から危害をもたらすインシデント結果に至る一連の事象を指す。危害の連鎖は AI 危害の特定において本質的な機能を果たす。AI 危害が存在するためには、AI システムの動作と発生した危害との間に直接的な関連性がなければならない。

この枠組みは、AIID から収集した 200 件以上のインシデントに適用され、インシデントタイプ、危害カテゴリー、その他の変数に基づいて注釈が付けられ分類された。[7] この分析プロセスと、多数のインシデントに対する詳細な調査により、危害の連鎖はインシデントの展開を形作る様々な力、すなわち危害メカニズムによって特徴づけられることが明らかになった。 人間と AI の双方のエラーが役割を果たした一方で、AI システムの意図的な悪用も影響し、時には特定の展開環境への AI 統合自体が単独で有害であった。上記の 6 つの被害メカニズムは、被害連鎖におけるこうした反復的な事象パターンの分析から導出されたものである。

第二段階では、AIID から無作為に抽出した 200 件のインシデントを分類することで、導出されたメカニズムの妥当性確認を行った。これには二つの理由があった。第一に、AI 被害枠組みが当初適用されたインシデントサンプルは無作為抽出ではなく、当時の AIID 内のインシデントの代表ではなかった。 次に、2023 年に第一段階を開始して以来、AIID のインシデント数は 50%以上増加していた。したがって、現在の技術革新レベルをより正確に反映する新たなインシデント群に対しても、これらのメカニズムが適用可能かつ関連性を保つかを妥当性確認することが不可欠であった。この妥当性確認結果は附属書の図 1 に示す。

制限事項

あらゆる枠組みやモデルは、必然的に現実を抽象化・簡略化した表現である。現実世界では、本報告書で示されるほど危害メカニズムが明確でない場合が多く、複数のメカニズムが同時に作用し得る。AI システムへの攻撃は、悪用を可能にするために実行されることが多い。モデルの性能問題と人間の監視ミスは同時に発生する。また、危害を意図して設計されたシステムが故障し、意図しない過剰な危害を引き起こすこともある。

意図性を二値で表現することは、介入対象となる異なる主体を区別するのに有用だ。しかし現実には意図性は連続体である。AI 利用の真に意図せず予見不可能な結果として生じるインシデントもあれば、明らかに意図的なインシデントもある。だが多くのインシデントはその中間にある。開発者や展開者が潜在的影響を考慮する怠慢、あるいは容易に予見可能な危害を無謀に無視した結果として生じるのだ。 この流動性により、設計による危害と統合上の危害の区別を困難にする境界事例が生じる。これらの事例を確実に判断するには、開発者や展開者の意思決定やガバナンス プロセスに関する洞察が必要であるが、通常は入手不可能だ。したがって、反証がない限り、本報告書は境界事例を分類する際、危害が意図的ではなかったと仮定する。今後の研究では、意図性をより細分化した表現を通じてこの限界に対処できるかもしれない。

同様に、情報の不足は設計による危害と誤用との区別を妨げる。外部観察者は通常、危害に関与したシステム内の基盤となる AI モデルを識別できないため、既存の AI モデルが誤用されたのか、それともこの目的のために設計されたものなのかが不明確な場合が多い。[8] それでも両カテゴリーの区別は重要である。なぜなら、AI バリューチェーン上のどの主体が危害を意図しているかによって、緩和が異なるからである。モデル開発者は設計による危害を引き起こし、AI システムの利用者は誤用による危害を引き起こす。[*] 両者に重なりがある場合でも、分離することで各メカニズムに効果的に対処する適切な緩和を識別できる。

最後に、データソースには二つの限界がある。AIID は現時点で最も包括的な AI インシデントと被害の集積だが、データベース内のインシデント類型の分布は必ずしも現実世界の頻度を反映していない。 このデータベースは報道機関の取材に依存しているため、メディア生態系の慣行やバイアスが見られる。そのため、注目を集めやすいインシデントや現在の社会的議論に関連するインシデントが過大評価されがちだ。これは、統合による被害のような目立たないメカニズムが、特に生成的 AI システムに関連する最近の社会的議論を牽引してきた人間の誤用事例に比べて過小評価されている可能性を示唆している。

開発者、展開者、利用者の区別は必ずしも明確ではなく、一つの事業体が複数の役割を担う場合もある。例えば ChatGPT は OpenAI が開発・展開する AI システムである。チャットボットとやり取りする個人が利用者となる。 OpenAI が自社の言語モデル GPT-5 を基盤に、サードパーティ（例：保険会社）向けの顧客サービスチャットボットを構築する場合、開発者は依然として OpenAI だが、展開者は保険会社であり、同社の顧客がチャットボットのユーザーとなる。

最後に、AI による多くの危害は、個別の事例として現れないため、インシデントデータベースで容易に捕捉できない。AI のエネルギー消費がもたらす結果、教育における AI への過度の依存が及ぼす悪影響、あるいは AI コンパニオンが人間関係に与える悪影響などは、個別のインシデントとして現れることが稀な潜在的な害の一例に過ぎない。分析に値するものの、こうした種類の危害は本研究の範囲外である。

AI の危害メカニズム

意図的な危害

設計による危害

害を意図して設計された AI システムは、6 つの危害のメカニズムの中で最も明白な形態である。この場合、開発者は AI システムが本質的に有害なタスクを実行するよう、あるいは有害な方法で利用されるよう設計する。設計による危害のインシデントでは、開発者と利用者の危害を及ぼす意図は概ね一致しているが、以下の例が示すように、設計による危害のシステムの種類は多岐にわたる。

防衛や法執行のために開発された AI システムの中には、標的指定に用いられる AI 搭載情報分析システムや戦闘空間管理システム、あるいは AI 搭載ナビゲーション・コンピュータビジョン・終端誘導を備えた自律兵器システムなど、設計による危害の明らかな例が存在する。 これらの AI システムが使用される際、開発者と運用者の双方が意図した結果が危害であるため、危害が発生するために故障や誤用は発生する必要がない。軍隊は、武力紛争法に従って展開される場合、合法的な戦闘員に対してこれらのシステムを適切に使用できる。ウクライナとイスラエルが関与した最近の紛争では、危害を引き起こす能力を持つ AI 搭載システムが戦闘に展開されたと報告されている。[9]

しかし、設計による危害は法執行や防衛の文脈以外でも蔓延している。

ユーザーが悪意を持って非同意の性的画像（NCII）を作成できるディープフェイクアプリが氾濫している。AIID には数十件ものインシデントが記録されており、個別に説明するには多すぎるほどだ。[10] 被害は圧倒的に女性や少女に及んでおり、こうしたインシデントは急速に拡大するジェンダーに基づくデジタル暴力の実態を如実に示している。[11] これは新たな問題でもなければ、AI の新たな能力でもない（画像生成技術は 2017 年頃から存在する）。しかし AI による「ヌディファイ」アプリやポルノ動画生成ツールがオンラインで広く利用可能になって以来、ポルノコンテンツ関連のインシデントは急増している。[12]

意図的に有害なアルゴリズム設計にはより巧妙な形態も存在する。ネイバー、クーパン、アマゾン・インディアなどのオンラインマーケットプレイスは、アルゴリズム操作による不公正な競争慣行に関与したとして非難されている。[13] これらの企業は、自社プラットフォームを支えるレコメンデーションシステムや検索アルゴリズムを不正に操作し、自社製品やブランドを優遇することで市場シェアを拡大し、競合他社に経済的・財務的損害を与えたとされる。 オンライン市場プラットフォームとしての支配的立場を悪用し、自社のブランドをベンダーとして推進することは独占禁止法に違反する。プラットフォームの広範なオンラインリーチと顧客基盤を考慮すると、わずかな操作であっても、その全体的な影響と被害の規模は甚大となり得る。

設計による危害の緩和

設計による危害への対処法は、その危害が社会的に容認可能か必要かによって異なる。許容できない危害をもたらす用途では、AI システムの開発・展開を禁止する措置が有効である。

防衛や法執行機能など、設計による危害が容認されると判断される状況では、ガバナンスの目的は害を完全に防止することではなく、明確に定義され異議申し立て可能な枠組みの中で、必要最小限に抑えることである。機構の政策は、過剰な害や過失による使用・悪用を防ぐため、技術の責任ある展開を確保するのに役立つ。一般的に組織は、自律的または AI による意思決定・行動への依存が許容される状況と条件を明確に定義した AI ガバナンス原則を確立すべきである。 役割と責任が明確に割り振られた堅固な説明責任の枠組みは、害をもたらす意思決定が透明かつ追跡可能であることを保証できる。十分な独立性と透明性を前提とした機関の監督

団体は、これらの政策や枠組みの違反が発生した、あるいは疑われる場合に、AI 支援の意思決定と実行された行動を調査・監査する認可を持つべきである。

AI の悪用

害を及ぼすことを明確な目的として開発されていない AI システムも、その目的のために悪用される可能性がある。設計段階での危害（開発者と利用者の双方に害意が存在する）と比較すると、AI の悪用では害意は AI システムの利用者または運用者のみに存在する。 なお、AI モデルは悪意のない目的でも悪用され得る。例えば宿題を AI に代行させる場合だ。[14] これによりユーザーが意図せず自らを害する可能性はある（自身の学習に悪影響を与えるなど）。しかし本節では、意図的な危害を伴う悪意ある悪用のみを扱う。(15)

専門的なアルゴリズムシステムも汎用 AI モデルも悪意ある悪用を受けやすいが、後者に関わるインシデントは AIID においてより一般的である。

大規模言語モデルやテキストから画像を生成するシステムを含む汎用 AI は、多様なタスクを高い精度で処理できるため、様々な目的での悪用が特に容易である。

例えば 2023 年には、オンライン掲示板 4chan のユーザーが ElevenLabs の音声合成 AI モデルを用いて有名人の憎悪的・暴力的な声のなりすましを作成した。[16] さらに最近では、マイクロソフトと OpenAI が北朝鮮、イラン、ロシア、中国の国家支援ハッカーが ChatGPT を悪用し、防衛・サイバーセキュリティ・暗号通貨分野を標的としたフィッシングやソーシャルエンジニアリング攻撃を行った事例を報告している。[17] 他の調査 では、ChatGPT がサイバー犯罪者によってマルウェアやその他の悪意のあるソフトウェアの作成に悪用されていたことが明らかになった。[18]

特定の単一目的のために設計された専門的な AI システムも、有害な悪用が可能だ。オンライン検索結果のランキングは懸念すべき事例である。完全または極めて高い自律性を示す AI システムを用いた検索結果のランキング操作は、悪意ある主体によって特に効果的に悪用され、有害で悪質な結果をもたらす。

例えば、反ユダヤ主義のオンライン集団が、車輪付き携帯オーブンの画像に「ユダヤ人のベビーカー」というラベルを付け加えた事例がある。[19] その結果、ユーザーが「ユダヤ人のベビーカー」という用語で検索すると、Google のアルゴリズムがオーブンの画像を検索結果の上位に表示した。 これは画像検索アルゴリズムの機能を直接悪用したもので、検索クエリの単語とウェブページ上の画像に付随する単語を照合する仕組みを利用している。この戦略が特に成功したのは、検索語に関連する「データの空白」があったためだ。実在しない製品「ユダヤ人のベビーカー」に関する検索結果は、攻撃的な画像のみであり、それがアルゴリズムによって上位表示されたのである。(20)

悪意あるユーザーは同様の組織的活動を行い、コンテンツモデレーションアルゴリズムを誘発してマイノリティのクリエイターの SNS 投稿を削除させる「敵対的通報」と呼ばれる戦術を展開している。[21] ソーシャルメディアサイトでは、投稿が実際にポリシー違反か否かに関わらず、一定数のユーザーからフラグが立てられるとコンテンツが自動削除される場合がある。この特性を利用し、右翼の荒らし集団は TikTok 上で少数派グループに属するインフルエンサーの投稿を戦略的に報告し、プラットフォームのコンテンツモデレーションアルゴリズムを起動させている。[22] TikTok の異議申し立てと審査プロセスで動画がコミュニティガイドラインに違反していないと判断された場合でも、投稿がフラグ付けされる頻度が高いほど罰則は厳しくなり、コンテンツ削除からアカウント削除まで及ぶ。こうして自動化されたコンテンツモデレーションシステムは、オンライン上でマイノリティコミュニティを効果的に検閲するために悪用されている。

悪用リスクに対する技術的緩和の限界

生成的 AI モデルの開発者は、有害なコンテンツの生成を制限するため、モデル出力を制御する措置を講じることができる。[23] 特に高性能なモデルへのアクセスを制限するリスクベースのリリース戦略は、悪用リスクへの対応をさらに支援する。[24] モデルの悪用傾向を評価するには、特定の悪意あるタスクを実行できるか（実現可能性）、実行可能であればその精度（性能）を評価する必要がある。[25] レッドチームは、様々な分野における悪用可能性を明らかにし、追加の安全対策が必要な箇所を浮き彫りにする手法として広く普及している。[26] 性能評価にはベンチマーク評価や実験を含め、既存のタスク実行手段と比較したモデルの限界効用に焦点を当てるべきである。[27]

包括的な安全対策を導入することは極めて困難である。開発者がモデルのあらゆる潜在的な（誤）使用を予測することは難しいからだ。最も重要なのは、モデルレベルでのこうした介入は、モデルの性能を低下させることなく悪用を必ずしも防止できないことだ。これは悪用と利用のトレードオフとして知られる。[28] AI モデルは悪意を理解する文脈を欠いているため、例えばフィッシングメールの作成を防ぐガードレールは、他のメールの作成も阻止する可能性が高い。コード作成においても同様だ。マルウェアを防止するガードレールは、無害なコンピュータプログラムの品質を低下させる可能性がある。現状の技術水準では、悪用不可能な AI システムを構築することは、非悪意的な目的ではほとんど役に立たないシステムを構築することを意味することが多い。AI 開発者や展開者がモデルの悪用を防ぐために講じられる他の手段はあるものの、技術的な修正だけでは悪用リスクを排除することはできない。

AI システムへの攻撃

AI システムへのサイバー攻撃からも被害が生じる。[*] 悪用メカニズムと同様、ここでも AI システムの開発者や展開者は害を意図していない。有害な意図は攻撃者側にある。サイバーセキュリティ分野では攻撃を三つのグループに分類する：機密性攻撃、完全性攻撃、可用性攻撃である。[29] 機密性攻撃は機密情報の抽出を、完全性攻撃はモデルの性能低下を、可用性攻撃はモデル全体の機能停止を目的とする。近年、生成モデルの安全対策を回避する新たな攻撃手法も出現している。[30] このような AI システムのセキュリティ脆弱性の悪用は、有害な結果を招く可能性がある。さらに、単独で引き起こす害に加え、AI システムへの攻撃はモデルの悪用を可能にする。

---

[*] AI モデルに対する敵対的攻撃は、AI ライフサイクルのあらゆる段階で発生し得る。AIID における関連インシデントは展開済みシステムへの攻撃に起因するため、本節では展開後の攻撃のみを扱う。

AI システムのセキュリティ脆弱性に関する証拠は豊富にあるが、AIID に記録された攻撃の大半は依然として実験環境下で発生しており、現実世界への被害には至っていない。例えばセキュリティ研究者は GitHub Copilot の脆弱性を発見しており、攻撃者が Copilot の応答を改ざんしたり開発者のデータを漏洩させたりする可能性があった（機密性・完全性攻撃）。[31] 実験により、テスラの自動運転システムの欠陥を悪用すれば、車両を加速させて対向車線へ逸脱させられることが示された（完全性攻撃）。[32] さらに調査では、ChatGPT に対する発散攻撃により、電話番号やメールアドレス、住所などの個人を特定できる情報を含む訓練データの漏洩を強制できることが判明した（機密性攻撃）。[33]

AIID によるインシデントは、実際のところ AI システムへの攻撃が、生成的 AI モデルの安全対策を回避するために行われることが多いことを示している。この手法は「脱獄」と呼ばれ、ユーザーがテキストプロンプトを考案し、AI モデルにポリシー違反の動作をさせるプロンプト・インジェクション攻撃に依存している。[34] プロンプト・インジェクション攻撃により、ユーザーは ChatGPT リリース直後にガードレールを回避し、差別的・暴力的なコンテンツを生成したり、犯罪行為の手順を提供したりした。[35] モデルが広範な安全性テストやレッドチームテストを経た後も、プロンプト・インジェクシ

ョン攻撃はガードレールを迂回する有力な手法であり続けている。マルウェア作成に大規模言語モデルを利用するハッカーは、OpenAI の GPT-4 や Anthropic の Claude など様々なモデルに対し、数十種類のプロンプト戦略を駆使して、フィッシングメールの作成、詐欺サイトの構築、マルウェア生成をモデルに実行させている。[36]

AI セキュリティは従来のソフトウェアセキュリティとは異なる

AI の脆弱性に対処し、より安全なモデルを構築することは特に困難である。なぜなら、確立されたサイバーセキュリティの実践や技術の多くが、機械学習モデルにそのまま適用できないからだ。[37] 機械学習モデルの確率的かつデータ依存的な性質は、従来のサイバーセキュリティにおける標準的な「パッチ適用と防御」モデルに本質的に不向きである。既知の脆弱性に対する効果的なパッチが存在しない場合や、実装コストが高すぎる場合がある。また、パッチ適用自体が新たなセキュリティ問題や性能上のトレードオフを引き起こし、修正の正当性を損なうほど深刻な影響を与える可能性もある。防御策は汎用性に欠けることも多く、特定の攻撃に対する対策が、同種の攻撃手法からモデルを保護するとは限らない。[38]

攻撃手段の多様性と既存防御策の信頼性不足を考慮すると、AI モデルの開発者・展開者はシステムレベルのアプローチでセキュリティを確保すべきである。[39] つまり、AI システムの一部が侵害される可能性を前提に展開し、冗長性を組み込んでシステムのレジリエンスを高める必要がある。言い換えれば、重要なモデル出力（決定、予測、行動）は単一の入力に依存すべきではなく、複数の測定値とアセスメントを組み合わせることで連鎖的な障害を防ぎ、個々のモデル要素への攻撃から防御する必要がある。

意図しない危害

AI の失敗

AIID における最も顕著な危害メカニズムは AI の失敗である。AI の失敗とは、システムが誤動作、機能不全、性能低下、バイアスを持つ出力を行う状況を広く包含する概念として定義される。この挙動は開発者、展開者、利用者にとって一般的に予期せぬものであり、以下に示す例が示すように、高リスクな状況で AI システムが使用された場合、壊滅的な結果を招きうる。

アルゴリズムバイアス議論の試金石となった論争として、ProPublica は 2016 年に再犯リスク予測アルゴリズム「COMPAS」を評価し、人種グループ間で性能に差異があることを発見した。[40] このアルゴリズムは全米で量刑決定や保釈前審問に用いられ、被告の刑罰の重さだけでなく、保釈金額や保釈前拘留の決定にも影響を与えていた。

調査によれば、黒人被告は白人被告のほぼ 2 倍の頻度で誤って再犯リスクが高いと判定され、一方白人被告は誤って低リスクと分類される確率が 2 倍高かった。このアルゴリズムが裁判官の被告アセスメントを導くため、裁判結果に具体的な影響を及ぼしていた。実際、ある被告の量刑が控訴審で減軽された後、裁判官は COMPAS が割り当てたリスクスコアによってより重い量刑に誘導されたことを認めた。プロパブリカの分析対象は、2013 年から 2014 年にかけて COMPAS リスクスコアを割り当てられた 1 万 8 千人以上に及んだ。

この 2 年間で、このアルゴリズムは黒人被告に対して過剰な保釈金額や長期刑を設定し、法執行機関や司法制度における歴史的な人種的バイアスを実質的に永続化・悪化させることで、数千件の不当な判決に関与した可能性が高い。[41]

司法制度以外にも、欠陥のある AI システムは医療分野でも展開されている。

一例がアーカンソー州のメディケイド免除プログラムで導入されたアルゴリズムだ。このプログラムは障害者向け在宅介護を提供しており、アルゴリズムは健康評価質問票の回答に基づき、受益者に割り当てる介護時間数を決定していた。多くの取得者の健康状態や介護ニーズは前年度のアセスメントから変化がなかったにもかかわらず、アルゴリズム導入により介護時間が大幅に削減された。 裁判でアルゴリズムが精査された結果、糖尿病や脳性麻痺といった重大な健康問題の処理に誤りがあり、プログラム受益者の 19%以上に誤った計算が行われていたことが明らかになった。[42] アルゴリズムの決定に対する苦情が継続していたにもかかわらず、展開機関も AI ベンダーも、訴訟によってソフトウェアの検査を迫られるまで誤りを検知できなかった。これは、アルゴリズムの決定によって不利益を被った者にとって、その決定を実質的に争う機会がいかに重要かを示している。[43]

AIID における AI の失敗例は他にも数多く存在する。生成的 AI システムはユーザーを脅迫し、偏った画像を生成し、誤情報を拡散した。[44]LinkedIn の検索アルゴリズムは、ある時期に男性のプロフィールを女性のプロフィールより優先していたようだ。[45]AI 搭載の運転支援技術は数多くの事故を引き起こし、中には死亡事故も含まれる。[46] また、誤った顔認識技術による不当逮捕のインシデントがデータベースに少なくとも 11 件記録されている。[47]

これらの AI 故障インシデントの発生は、少なくとも次の二つのうちいずれかが真実であることを示している（開発者の指示通りに使用した場合を前提とする）：既存の評価、リスクマネジメント、保証の実践が、これらの AI システムを安全に展開できることを保証するのにまだ十分でないか、あるいはこれらの実践が AI 開発者や展開者によって不十分に行われているかのいずれかだ。緩和策は、AI 故障の可能性を減らすために、この両方の欠点を対処する必要がある。

標準や規制がない状況では、救済手続きによって AI 故障による被害の影響を緩和できる

AI の精度と信頼性を高めることは、モデル性能の問題を確実に検知できる（そして性能自体を評価できる）テスト・評価プロトコルの開発と同様、継続的な研究分野である。[48]AI 標準の領域もまだ発展途上であり、開発者が AI システムの安全性・正確性・信頼性を確保するために頼れる合意された技術や手法は、まだ多く存在しない。 仮に存在したとしても、規制要件がなければ開発者や展開者は AI 安全対策の実施を自由に選択できる。実環境での導入前テストや導入後監査といったコストのかかる対策は、義務付けられない限り実施される可能性は低い。

強力な規制や厳格な標準があっても、ある程度の AI の失敗は避けられないだろう。したがって、失敗が発生した際の被害を軽減するために展開者が何ができるかを検討する価値がある。例えば、前述のインシデントは、エラーやバイアス、その他の失敗が発生した際に、AI システムの決定を意味ある形で検証する機会が重要であることを示している。 司法制度は被害者への救済手段となり得るが、それは十分な資源を持つ者に限られる。AI 支援決定に対する人間の審査を含む苦情処理・救済手続きは、アルゴリズムによる被害への救済アクセスを強化し、AI システムの展開に伴い司法制度が過負荷になるのを防ぐため、導入者レベルで実施されるべきである。

人間の監視の失敗

機微な状況では、AI システムは人間と機械のチームの一環として展開されることが多い。アルゴリズムによる意思決定支援システムでは、AI システムの推奨事項を考慮しつつ、最終決定権を人間オペレーターが保持する必要がある。理想的には、AI システムの出力に対する人間の監視は、異常な動作、バイアス、性能問題を検知し、発生した場合は意思決定プロセスでそれらを排除することを保証すべきである。 しかし実際には、人間の監督者は問題を検知できないだけでなく、必要な場面で AI システムの推奨を覆すこともできないことが多い。

この顕著な例が、欠陥のある音声認識システムの結果に基づいて英国政府が数千件の移民ビザを取り消した事件だ。[49] 英国への移民は、国際試験機関 ETS が実施する「国際コミュニケーション英語テスト（TOEIC）」に合格することで英語能力を証明しなければならない。 ETS は代理受験者による不正を検知するため、音声認識ソフトウェアを展開し、複数の試験録音に同一の声が含まれていないかを判定した。不正が疑われる場合、ETS 職員 2 名が合意して初めて試験結果が無効と判定される仕組みだった。展開から 3 年間で、TOEIC 試験 の録音の 97%が音声認識 AI によって疑わしいとフラグ付けされた。 この明らかな異常にもかかわらず、審査員はこれらの試験の半数以上を無効（その他は全て疑わしい）と分類し、無効試験のリストを英国政府に提出した。この結果に基づき、英国の政策立案者は告発された受験者のビザを取り消し、国外退去手続きを開始した。[(50)]

この事例における監視の失敗は、人間が AI システムと関わる際の心理的傾向に起因している可能性が高い。人間による審査員が専門知識を有していたか、音声類似性を評価する訓練を受けていたかは不明だが、彼らの判断は二つの認知バイアスに影響されたと考えられる。第一に自動化バイアスとは、機械が生成した出力を実際以上に客観的で真実であると受け入れる傾向である。 二つ目はアンカリングバイアスであり、人間が AI システムの予測を「アンカー」として利用し、自身の判断でそこから大きく逸脱しない傾向を指す。[51] バイアスは個々の判断からは検知しにくいが、大量の判断を分析すると明らかになる。本件では明らかに、事後的な大規模アセスメントが行われていなかった。

認知バイアスに加え、人間意思決定者がアルゴリズムの推奨を覆すことを妨げる構造的要因も存在する。ナルクスケアの事例がこれを示している。ナルクスケアは全国処方薬データベース上に構築されたアルゴリズムツールで、医師や薬剤師にデータ分析と医療管理機能を提供する。 その一つが、オピオイド乱用・過剰摂取リスクを自動算出する患者リスクスコアである。[52] ツールプロバイダは「治療判断の自動化ではなく参考情報として活用すべき」と強調するが、医療従事者は予測スコアを覆す上で重大な障壁に直面している。多くの州では医師・薬剤師が規制薬物処方前に本プログラムを参照することが法的に義務付けられており、一部州では令状なしに法執行機関がデータにアクセスし医師と患者双方を起訴できる。 結果として、医師は過剰処方と見なされることをますます恐れるようになった。免許剥奪を恐れる処方者は、極端な慎重さに走り、薬剤決定時にリスクスコアに過度に依存する。これにより、正当な必要性がある患者でさえ処方量が全体的に減少している。[53] 監視体制とアルゴリズムツールの強制使用が相まって、医師の患者ケアに対する主体性が損なわれているのだ。 慢性疼痛患者や複数の重篤な疾患を抱える患者にとって、これは必要な薬の入手困難、痛みの増大、生活の質の低下を招いている。

自動運転や運転支援技術を搭載した車両の運用においても、人間の監視機能の欠如が頻発している。部分自動運転においては、人間は「ループ内」ではなく「ループ外」に位置づけられる。この微妙な差異は、意思決定・行動プロセスにおける人間の関与形態が能動的から受動的へと大きく転換したことを意味する。 部分自動運転車における「オン・ザ・ループ」とは、システムが運転者の監視下で自律的に動作し、理論上は運転者がいつでも制御を引き継げる状態を指す。しかし人間は、長時間にわたりシステムを警戒しつつ受動的に監視する能力に欠けている[54]。その結果、介入すべき状況を見逃したり反応が遅れたりして事故を引き起こすことがある。 この人間と機械の連携の失敗は、AIID（自動運転事故データベース）に記録された AI 搭載車関連の多数のインシデントで確認できる。[55] 主要な自動運転車製造事業者の調査を担当する米国政府機関である国家道路交通安全局（NHTSA）は、この問題を「自動化過信」と名付けた。これはシステム能力への過剰な信頼と、監視タスクから注意を逸らしやすい人間の性質が組み合わさった現象である。[56] 当局の調査結果によれば、自動化過信は少なくとも部分的に製造事業者の不適切な設計選択に起因している。AI 搭載運転支援技術が関与した交通事故の調査では、衝突のわずか 1 秒未満前に車両の自動操縦機能が制御を人間運転者に引き渡す事例が多数確認された。これにより運転者は事故を回避するために車両を完全制御する十分な反応時間が与えられなかった。[57] 人

間の能力や限界（注意持続時間や反応時間など）を考慮した人間と機械のインターフェースを改善することは、状況に応じて人間が AI システムの制御を引き継げるようにするために極めて重要だ。

人間と AI のチームにおける認知バイアスの緩和課題

認知バイアス、制度的環境、AI システムの設計は、人間の AI システム監視の成否を左右する重要な要素である。[58] 人間の監視が AI の危害を効果的に緩和するためには、これらの側面を慎重に検討し対処する必要がある。

しかし、人間と AI の相互作用における認知バイアスの軽減は未解決の課題である。なぜなら、バイアスを緩和すると期待される多くの対策が、逆にバイアスを悪化させる可能性があるからだ。AI の説明可能性機能は、過度の依存傾向を緩和しシステムの出力に対する信頼構築に寄与すると考えられているが、場合によっては認知バイアスを悪化させることが判明している。[59] AI リテラシーについても同様の効果が確認されている。AI リテラシーが低いほど AI の推奨への依存度が高まる一方で、AI リテラシーが高いほどアルゴリズム回避（AI の推奨に対する過度な不信感）が生じ、結果として意思決定の精度が低下する傾向があった。[60]

認知バイアスの緩和手法を明らかにし、意思決定者の主体性を守り、人間と AI のチームへの信頼を構築するためには、人間工学と人間コンピュータ相互作用に関するさらなる実験的研究が必要である。

統合的な危害

最後に、不適切な文脈で使用される AI システムは、意図通りに機能し、有害な意図がなく、攻撃や悪用もされていない場合でも危害を引き起こす可能性がある。こうしたケースでは、 AI が展開環境と統合された結果として、意図しない形で危害が生じる。 他の危害メカニズムと比べ、このケースの引き金は故障や悪意あるユーザーではなく、本来機能するシステムが不適切な文脈に展開されることにある。AI システムが有害に設計される場合と同様に、統合的な危害はシステムの展開が直接的な原因となるが、それは明示的な目的ではなく意図しない副作用として現れる。

単独で見れば、AI ツールは一見無害な機能を果たすことがある。例えばウェブページの順位付けによるオンライン検索の支援や、店舗への来客数の予測などだ。非常に成功している場合すらあるが、展開環境への統合によって意図しない副作用が生じうる。これは全ての AI ツールが既存システムに組み込まれるためだ。店舗の来客予測ツールは職場環境の一部となり、既存の業務フローに統合される。 オンライン検索エンジンは、我々のオンライン情報環境という文脈の中に位置づけられる。AI システムをこうした文脈に組み込むことで、環境内に変化が生じ、潜在的な害をもたらす可能性がある。以下の事例は、文脈を考慮しなかった結果、害が生じた事例を示している。

過去 20 年間で、オンライン検索は人々が健康関連情報にアクセスする主要な手段となった。[61] ページランクアルゴリズムで駆動されるオンライン検索エンジンは、検索結果の表示順序を通じて、インターネットユーザーが発見し消費する情報に直接影響を与える。コロナウイルスパンデミック中、ワクチン反対の誤情報は、オンライン検索エンジンを駆動するそれらのアルゴリズムによって増幅された。アマゾンのアルゴリズム監査では、同プラットフォームがワクチン陰謀論を扱う書籍など誤情報を拡散する多様な商品をホストしているだけでなく、その検索エンジンがそうした商品を優遇するランキングバイアスを示していることが明らかになった。ユーザーがワクチン関連の一般的な用語を検索すると、正確な情報や反論を提供する商品よりも、ワクチンに関する誤情報を流す商品が上位に表示されるのだ。[63] 監査はさらに、プラットフォームのレコメンデーションアルゴリズムによるフィルターバブル効果を実証した。誤情報を提供する商品と接触したユーザーは、Amazon のホームページやおすすめ商品の中で誤情報に遭遇する可能性が高かった。こ

のように、顧客に人気のある関連商品を推奨することを意図した検索・レコメンデーションアルゴリズムの通常の機能は、Amazon の顧客に誤情報を広めるという意図しない副作用をもたらしたのである。

アルゴリズム設計においてステークホルダーのニーズを一方的に考慮した結果、意図せぬ結果が生じることもある。スターバックスが全店舗に展開した勤務スケジュールアルゴリズムがこれに該当する。この AI システムは予測される店舗客数に基づき従業員のシフト配分を最適化するために開発された。しかし展開の結果、常に変動するシフトスケジュール（多くの場合、従業員への通知は直前に行われた）と週ごとの労働時間の劇的な変動が生じた。[64] このツールは小売チェーンのコスト削減を優先し、従業員が予測可能な勤務スケジュールと収入を求めるニーズを無視した。意図的な害ではなかったーースターバックスは従業員のニーズを考慮しなかったが、彼らに苦難をもたらすためにシステムを展開したとは考えにくいーーが、このシステムの使用は従業員の経済的安定性と、仕事以外の生活を計画・管理する能力に深刻な影響を与えた。[65]

最後に、AI システムの展開はワークフローを混乱させ資源を振り向けることで機会費用を生む可能性がある。その一例が、2016 年にシカゴ警察が展開した音響式銃声検知システム「ショットスポッター」だ。2024 年の評価では、展開後 6 年間でこの AI システムが 1 日あたり約 70 件の出動要請を発生させ、導入前の 2 倍に増加したことが判明した。 この警官リソースの需要増加は、911 通報への対応時間に影響を与えた。警官は 911 通報への出動が遅くなり、現場到着も遅れた。さらに犯人を逮捕する可能性も低下した。したがって、銃声検知システムの展開は、シカゴ市民の緊急事態への警察対応の有効性を低下させたのである。[66]

アルゴリズム影響アセスメントは統合的害悪のリスクを緩和できる

統合的な危害のリスクを低減するには、AI システムの展開前に潜在的な影響を予測し、それに応じてアルゴリズム設計を調整する必要がある。EU AI 法で義務付けられている基本的権利影響評価のような事前 AI 影響評価は、これを達成するための措置の一例だ。 AI 影響評価の目的は、AI のユースケースと使用環境の分析、および展開の影響を受ける人々との協議を通じて、AI システムの使用から生じうる幅広い危害を明らかにし評価することにある。アルゴリズム影響評価は、導入者および影響を受ける関係者に、AI システムの採用を決定する前にその効果を評価する機会を与える。[67] これを効果的に行うには、AI システムの動作原理に関する知識と、使用状況の文脈に対する深い理解の両方が必要である。 多くの場合、これは多様な関係者を巻き込み、影響を受けるグループからの懸念や視点を調査することを意味する。影響評価はユースケースごとに実施され、各展開状況と目的に合わせて調整されるべきである。その結果は、アルゴリズム設計の選択、リスク緩和戦略、そして最終的には展開決定に反映されるべきである。

議論

AI の有害メカニズムを調査・分類することは貴重な知見をもたらす。第一に、AI の使用が意図的・非意図的に害を及ぼす多様な方法を明らかにする。AI システムが誤動作したり悪用されたりしなくても有害となり得ることを認識することは、リスク緩和アプローチを広げる重要な第一歩だ。 特に統合的危害は、AI 展開の多面的な影響を予測するのが難しいため、見過ごされがちなメカニズムだ。本報告書は、様々な AI 危害メカニズムを識別・説明することで、リスク緩和策が危害につながる全ての潜在的な経路に対処することを目指す。

第二に、AI の危害のメカニズムの広範さを考慮すると、同様に多様な緩和アプローチの必要性が明らかになる。画一的なアプローチは通用しない。代わりに、AI バリューチェーンに関わる多様な主体（開発者、展開者、政策立案者など）が、様々な社会技術的緩和策を幅広く講じることで、異なる害のメカニズムが発生するリスクを低減できる。 本報告書で論じた通り、AI モデルへの技術的修正だけでは AI による危害を防げない。対策は、テスト・評価プロトコルの実施から、AI 導入における参加型プロセスまで多岐にわたる必要がある。

第三に、様々な危害メカニズムを分析すると、AI システムの危害発生傾向は必ずしもその能力と結びついているわけではないことが明らかになる。実際、本文中の多くの事例は、特殊化された単一目的の AI から危害が生じたケースを示しており、規模や能力、高度さ（しばしば訓練に必要な計算量で間接的に測定される）にかかわらず、あらゆる種類の AI システムから危害が生じ得ることを示している。 害のリスクをより強く決定づけるのは、AI の設計、テスト、展開における文脈、方法、ガバナンスである。

最後に、現実世界での被害事例報告は AI リスクを学ぶ上で貴重な資源であり、より良い緩和策や的を絞った政策介入を設計するための本質的な知見を提供する。CSET の AI 被害枠組みと分類体系、MIT の AI インシデントトラッカーなど、既存の事例データをよりアクセスしやすく分析可能にし、政策知見を導き出すための多くの取り組みが進行中だ。[68] しかし、本報告書の限界事項で一部説明されているデータ上の制約は、厳密な分析における重要な障壁であり続けている。正式なインシデント報告制度は、関与した AI システムに関する詳細情報を収集することで、既存の追跡努力の欠点を克服できる。安全かつ責任ある AI 展開を推進する政策立案者にとって、このような義務的なインシデント報告制度の確立は優先課題であるべきである。

結論

AI の能力が向上し、潜在的な活用事例が多様化するにつれ、公共部門と民間部門の双方が、生産性向上などの利益を実現するため、より強力な AI 展開を推進している。しかし、AI を生活のあらゆる側面に組み込む前に、それによる危害リスクをより深く理解することが極めて重要だ。AI の広範な展開と利用は、既に数百件の記録された危害インシデントを引き起こしている。 将来の危害をより効果的に防止するには、実際に危害がどのように発生するかをより深く理解する必要がある。この目的のために、本分析では AI インシデントデータベースの事例報告を活用し、AI の使用が有害な結果につながる 6 つの主要な危害メカニズムを特定した（表 3）。

表 3：6 つの AI 危害メカニズム

意図的な危害	意図しない危害
● 設計による危害 ● AI の悪用 ● AI システムへの攻撃	● AI の失敗 ● 人間の監視の失敗 ● 統合的な危害

これらのメカニズムに関連するインシデントを検証した結果、将来の被害発生を低減するための政策努力を導くべき重要な示唆がいくつか明らかになった：

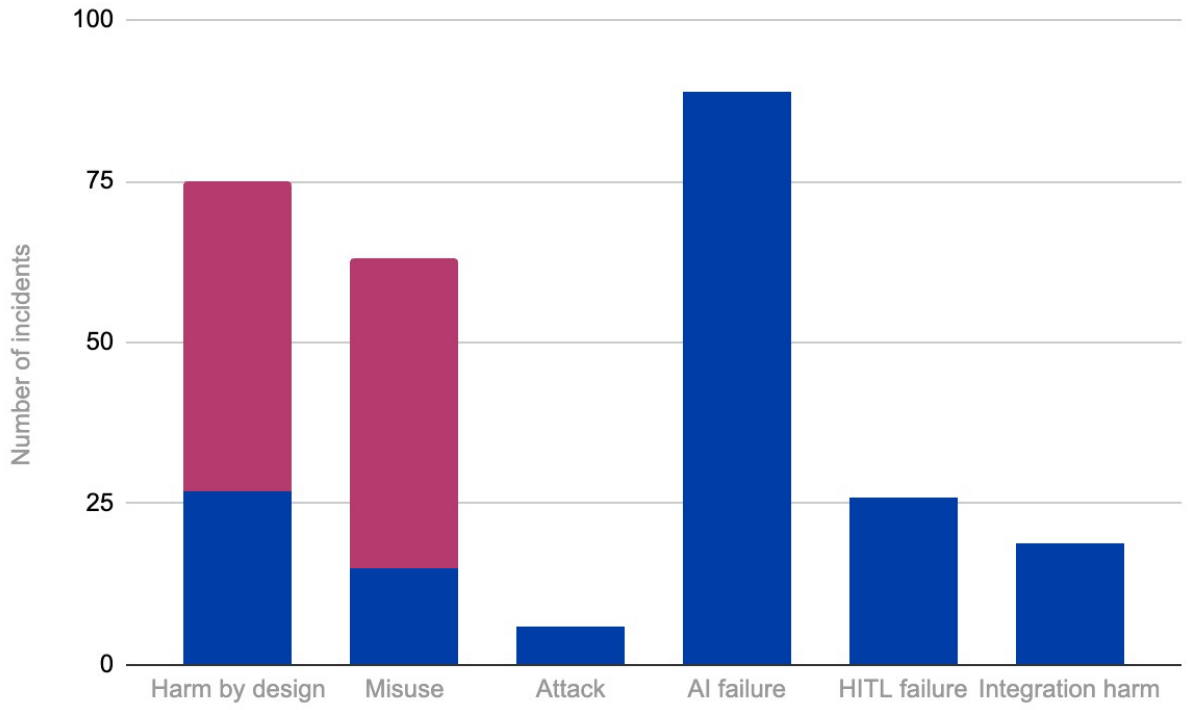
1. 危害軽減への画一的なアプローチは機能しない。本報告書が示す通り、危害に至る経路は多様であり、同様に多様な緩和戦略を必要とする。純粋に技術的なアプローチでは不十分であり、特に統合的な危害や人間の監視の失敗に対処する上で限界がある。

2. 計算能力で測られるモデルの能力は、危害発生の傾向を予測する上で不十分である。本報告書は、単一目的の AI システムが危害に関与した事例を数多く示している。高度な AI システムにリスク緩和策を集中させても、専門的な AI システムの無責任な設計・展開・使用から生じる現実的なリスクには対処できない。

3. AI がもたらすリスクを識別し対応する能力を高めるには、包括的なインシデント追跡が不可欠である。広範な社会技術的緩和策を実施すれば AI による被害発生を大幅に減らせるが、インシデントを完全に防ぐことはできない。 AI の革新が新たな機能と新たな故障モードを明らかにし、展開者が新たな利用ケースを設計し、悪意ある主体が AI システムを攻撃・悪用する新たな方法を見出すにつれ、新たな被害が発生する。技術革新に追いつくためには、インシデント報告からの効果的な学習によって可能となる、機敏な対応と緩和策の迅速な適応が必要である。

附属書

図 1：AI による危害メカニズムの分布



注：200 件の AI インシデントの無作為抽出サンプルに対する危害メカニズムの注釈。 数値の合計が 200 に達しないのは、単一のインシデントに複数のメカニズムが関与する場合があるためである。「設計による危害」と「誤用」は 2 本の棒グラフで示されている。青色の棒グラフは、使用された AI システムに関する情報が利用可能で、モデルが誤用されたか、危害を目的として開発されたかを判断するのに十分な、確認済みの事例を反映している。赤色の棒グラフは、この判断が不可能であった 48 件のインシデント、すなわち設計による危害または誤用の事例である可能性があるものを表している。

出典：AI インシデントデータベースのデータに基づく著者の分析。

図は AIID から無作為に抽出した 200 件のインシデントにおける危害メカニズムの分布を示す。AI の失敗が圧倒的に多い原因であり、サンプル中の 89 件がこのメカニズムに関連している。これは AI システムが（まだ）確実に遂行できない目的で展開され続け、重大な危害をもたらしていることを示唆している。

AI の失敗以外に、設計上の欠陥や誤用による被害事例の多さも注目に値する。しかし、個々のインシデントで使用された AI システムに関する知見が不足しているため、根本的な問題の理解は妨げられている。 これは、AI システムを構築するリソースを持つ主体が増加し、AI 搭載の「ヌディファイ」アプリやマルウェア生成生成的 AI といった有害モデルが普及しているためだろうか？ それとも、開発者の意図に反して AI システムが濫用されているのだろうか？ これらの疑問に答え、リスクの緩和に適切なリソースを配分するためには、インシデント追跡とデータ収集への投資が不可欠である。

概して、ほとんどのインシデントは意図しない危害メカニズムから発生する。これは意図的なメカニズムほど正確に危害の原因を特定できないため、緩和をより複雑にする。攻撃者や悪意あるユーザーといった動機ある主体を危害の源として識別できれば、具体的かつ標的を絞った緩和戦略を構築できる。 しかし意図しない危害メカニズム、特に

統合上の危害や人的監視の失敗に対しては、より多面的で複雑なアプローチが必要となる。多くの場合、より多くの関係者を巻き込むアプローチが求められるのである。

著者

ミア・ホフマンは CSET の研究員であり、AI ガバナンスを専門としている。

謝辞

筆者は、本報告書の初期ドラフトに対するコメントを寄せてくれたジョシュ・ゴールドスタイン、コリン・シェイ＝ブライマイヤー、ミナ・ナラヤナン、マドゥ・スリクマール、ダニー・アサートン、ならびにフィードバックと支援を提供してくれたオーウェン・ダニエルズに感謝する。

© 2025 センター・フォー・セキュリティ・アンド・エマージング・テクノロジー。本著作物はクリエイティブ・コモンズ表示-非営利 4.0 国際ライセンスの下で提供されている。

このライセンスのコピーを見るには、https://creativecommons.org/licenses/by-nc/4.0/ にアクセスせよ。

文書識別子: doi: 10.51593/20240052

脚注

[1] ショーン・マクレガー、「インシデントのカタログ化による現実世界での AI 失敗の再発防止：AI インシデントデータベース」、第 *33* 回人工知能の革新的応用に関する年次会議（*IAAI-21*）論文集 35 巻 17 号（2021 年）。

[2] 行政管理予算局、『イノベーション、ガバナンス、公共の信頼を通じた連邦政府の *AI* 活用加速』（ワシントン D.C.：大統領府、2025 年 4 月 3 日）、https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AIthrough-Innovation-Governance-and-Public-Trust.pdf; 欧州議会及び理事会規則（EU）2024/1689（2024 年 6 月 13 日発効）人工知能に関する調和規則を定めるもの（人工知能法）、OJ L, 2024/1689, https://eur-lex.europa.eu/eli/reg/2024/1689/oj; 「Anthropic の責任ある拡張方針」、Anthropic、2025 年 5 月 14 日、https://www.anthropic.com/rsp-updates。

[3] ヘザー・フレーズとミア・ホフマン、「AI の危害に構造を与える」（安全保障と新興技術センター、2023 年 7 月）、https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/。

[4] 「CSETv1 チャート」、AI インシデントデータベース、https://incidentdatabase.ai/taxonomy/csetv1/。

[5] 欧州委員会、「汎用 AI 行動規範」、https://digitalstrategy.ec.europa.eu/en/policies/contents-code-gpai。

[6] Frase and Hoffmann, 「AI による危害に構造を与える」

[7] 「CSETv1 チャート」、AI インシデントデータベース、https://incidentdatabase.ai/taxonomy/csetv1/。

[8] AI インシデントデータベースのインシデント 544、565、673 はこれをよく示している。

[9] カテリーナ・ステパネンコ「戦場における AI 革命はまだ到来していない：ロシアとウクライナの現行 AI ドローン開発の現状」（戦争研究所、2025 年 6 月 2 日） https://understandingwar.org/research/russia-ukraine/the-battlefield-ai-revolution-is-not-here-yet-thestatus-of-current-russian-and-ukrainian-ai-drone-efforts/; シーラ・フレンケル、ナタン・オデンハイマー「ガザ戦争におけるイスラエルの AI 実験が倫理的懸念を招く」『ニューヨーク・タイムズ』2025 年 4 月 25 日、https://www.nytimes.com/2025/04/25/technology/israel-gaza-ai.html; 「インシデント番号 672：ラベンダーAI システムがガザ攻撃を指揮、民間人犠牲率が高いと報告」人工知能インシデントデータベース、Responsible AI Collaborative、2024 年 4 月 3 日、https://incidentdatabase.ai/cite/672/。

[10] 例として、AI インシデントデータベースのインシデント番号 480、530、597、765、769、771、772、777、874、903 を参照せよ。

[11] Megan Hughes, 「テクノロジーが女性と少女への暴力を助長する仕組み」 (新興技術・安全保障センター, 2024 年 8 月), https://cetas.turing.ac.uk/publications/howtechnology-can-enable-violence-against-women-and-girls.

[12] マット・バージェス「ディープフェイクポルノは制御不能だ」『*Wired*』2023 年 10 月 16 日 https://www.wired.com/story/deepfake-porn-is-out-of-control/?_sp=7405a578-d28d-4413-9f8a8fbbb4387704.1740429425353.

[13] 「インシデント番号 191：韓国インターネットポータル大手ネイバー、自社サービスを優遇するためショッピング・動画検索アルゴリズムを操作」人工知能インシデントデータベース、Responsible AI Collaborative、2020 年 10 月 6 日、https://incidentdatabase.ai/cite/191/; ソナリ・ペドネカール「インシデント番号 435：クーパンが自社製

品を優遇するため検索アルゴリズムを調整した疑惑」人工知能インシデントデータベース、Responsible AI Collaborative、2025 年 7 月 4 日、https://incidentdatabase.ai/cite/435/; ソナリ・ペドネカール、「インシデント番号 437：アマゾン・インディアが自社製品を促進するため検索結果を不正操作した疑惑」、人工知能インシデントデータベース、Responsible AI Collaborative、2016 年 12 月 31 日、https://incidentdatabase.ai/cite/437/。

[14] Khoa Lam,「インシデント番号 339：学生による課題・試験不正行為へのオープンソース生成的モデルの悪用」, 人工知能インシデントデータベース, Responsible AI Collaborative, 2022 年 9 月 15 日, https://incidentdatabase.ai/cite/339/.

[15] Natalya Kosmyma, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein and Pattie Maes,「ChatGPT が脳に与える影響：AI アシスタントを用いた論文執筆タスクにおける認知負債の蓄積」, arXiv プレプリント arXiv:2506.08872, https://arxiv.org/pdf/2506.08872v1.

[16] ジャネット・シュワルツ、「インシデント番号 508：悪意ある目的で悪用された有名人のディープフェイク音声」、人工知能インシデントデータベース、Responsible AI Collaborative、2023 年 1 月 30 日、https://incidentdatabase.ai/cite/508/。

[17] ダニエル・アサートン、「インシデント番号 644：国家支援とされるハッカーが人工知能を用いたフィッシング攻撃をエスカレート」、人工知能インシデントデータベース、Responsible AI Collaborative、2024 年 2 月 18 日、https://incidentdatabase.ai/cite/644/。

[18] 「インシデント番号 736：LLM の闇市場がマルウェアとフィッシング詐欺を助長」、人工知能インシデントデータベース、Responsible AI Collaborative、2023 年 12 月 1 日、https://incidentdatabase.ai/cite/736/; ダニエル・アサートン、「インシデント番号 443：悪意のあるソフトウェア開発に悪用された ChatGPT」、人工知能インシデントデータベース、Responsible AI Collaborative、2022 年 12 月 21 日、https://incidentdatabase.ai/cite/443/。

[19] デヴォン・コルマー、「インシデント番号 1057：『ユダヤ人のベビーカー』が反ユダヤ主義的な Google 画像のプロバイダ、ヘイトスピーチキャンペーンの結果とされる」、人工知能インシデントデータベース、Responsible AI Collaborative、2017 年 8 月 15 日、https://incidentdatabase.ai/cite/1057/。

[20] Michael Golebiewski と Danah Boyd、「データの空白：欠落データが容易に悪用される場所」（Data & Society、2018 年 5 月 11 日）、https://datasociety.net/library/data-voids-where-missing-data-caneasily-be-exploited/。

[21] 別のタイプの敵対的報告については、例えば以下を参照のこと。シェルビー・グロスマン、クリストファー・ジャイルズ、シンシア・N・M、R・マイルズ・マケイン、ブレア・リード「新たな著作権トロール：Twitter ネットワーク がタンザニアの活動家を嫌がらせするために著作権侵害申し立てをどう利用したか」（スタンフォードインターネット観測所、2021 年 12 月 2 日）、https://cyber.fsi.stanford.edu/io/publication/new-copyright-trolls.

[22] Patrick Hall、「インシデント番号 133：オンラインの荒らしが、TikTok の自動コンテンツ報告システムを悪用して、周縁化されたクリエイターを識別的に差別したとされる事件」、人工知能インシデントデータベース、Responsible AI Collaborative、2020 年 12 月 15 日、https://incidentdatabase.ai/cite/133。

[23] Jessica Ji、Josh A. Goldstein、アンドルー・ローン、「大規模言語モデルの出力制御：入門編」（セキュリティ・新興技術センター、2023 年 12 月）、https://cset.georgetown.edu/publication/controlling-large-language-models-a-primer/。

[24] ジョン・ベイトマン、ダン・ベア、ステファニー・A・ベル、グレン・O・ブラウン他、「オープン対クローズドを超えて：基礎となる AI モデルのガバナンスに関する新たなコンセンサスと重要な課題」（カーネギー国際平和財団、2024 年 7 月 23 日） https://carnegieendowment.org/research/2024/07/beyond-open-vsclosed-emerging-consensus-and-key-questions-for-foundation-ai-model-governance?lang=en。

[25] ジョシュ・A・ゴールドスタインとギリッシュ・サストリー、「PPOu 枠組み：高度な AI システムの悪用可能性を評価する構造化されたアプローチ」、*AAAI/ACM* カンファレンス「*AI*、倫理、社会」論文集 7 巻 1 号（2024 年 10 月）：503–518 頁、https://doi.org/10.1609/aies.v7i1.31653.

[26] Jessica Ji,「AI レッドチームとは実際何を意味するのか？」, CSET ブログ, 2023 年 10 月 24 日, https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean/.

[27] トーマス・ウッドサイドとヘレン・トナー、「大規模言語モデルの評価」、CSET ブログ、2024 年 7 月 17 日、https://cset.georgetown.edu/article/evaluating-large-language-models/; ゴールドスタインとサストリー、「PPOu 枠組み」

[28] Markus Anderljung、Julian Hazell、Moritz von Knebel、「AI の悪用から社会を防御する：能力制限が正当化されるのはいつか？」、『AI & Society』40: 3841–3857 (2025) https://doi.org/10.1007/s00146-024-02130-8; アルヴィンド・ナラヤナンとサヤシュ・カプール、「AI の安全性はモデルの特性ではない」、『AI Snake Oil』、2024 年 3 月 12 日、https://www.aisnakeoil.com/p/ai-safety-is-not-amodel-property。

[29] アンドルー・ローン、「AI のハッキング：機械学習のサイバーセキュリティに関する政策立案者向け入門書」（セキュリティ・新興技術センター、2020 年 12 月）、https://cset.georgetown.edu/publication/hackingai/。

[30] Apostol Vassilev、Alina Oprea、Alie Fordyce、Hyrum Anderson、「敵対的機械学習：攻撃と緩和の分類と用語集」（ワシントン D.C.： 国立標準技術研究所、2024 年 1 月）、https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf。

[31] Marlon Fabiano、「Github [原文のまま] Copilot のゼロデイ」、gccybermonks、2024 年 7 月 27 日、https://gccybermonks.com/posts/github/、https://incidentdatabase.ai/reports/3996/ から取得。

[32] Patrick Howell O'Neill,「ハッカーがテスラを騙して時速 50 マイル加速させる方法」, *MIT Technology Review*, 2020 年 2 月 19 日, https://incidentdatabase.ai/reports/2316/; エヴァン・アッカーマン、「交差点に貼った 3 枚の小さなステッカーがテスラの自動運転を誤った車線へ逸らす原因となる」、IEEE Spectrum、2019 年 4 月 1 日、https://incidentdatabase.ai/reports/1518/ から取得。

[33] ミラド・ナスル、ニコラス・カーリーニ、ジョナサン・ハヤセ他、「（実稼働）言語モデルからのトレーニングデータのスケーラブルな抽出」、arXiv プレプリント arXiv: 2311.17035 (2023)、https://arxiv.org/abs/2311.17035。

[34] ダニエル・アサートン、「インシデント番号 473：初期テスターによるプロンプト・インジェクションで明らかになった Bing Chat の初期プロンプト」、人工知能インシデントデータベース、Responsible AI Collaborative、2023 年 2 月 8 日、https://incidentdatabase.ai/cite/473/; コリン・スタークウェザー「インシデント番号 677：ChatGPT と Perplexity が AI ボーイフレンドシナリオでコンテンツポリシー違反を誘導されたと報告」人工知能インシデントデータベース、Responsible AI Collaborative、2024 年 4 月 29 日、https://incidentdatabase.ai/cite/677。

[35] Khoa Lam,「インシデント番号 420：ユーザーが ChatGPT のコンテンツフィルターを容易に回避」,『人工知能インシデントデータベース』, Responsible AI Collaborative, 2022 年 11 月 30 日, https://incidentdatabase.ai/cite/420.

[36] 「インシデント番号 736：LLM の闇市場がマルウェアとフィッシング詐欺を助長」、『人工知能インシデントデータベース』、Responsible AI Collaborative、2023 年 12 月 1 日、https://incidentdatabase.ai/cite/736/; David Winder、「新しい AI 攻撃が Google Chrome のパスワードマネージャーを危険にさらす」、*Forbes*、2025 年 3 月 23 日、https://incidentdatabase.ai/reports/5011/。

[37] アンドルー・ローンとワイアット・ホフマン、「AI のセキュリティ確保：従来の脆弱性開示の適応方法」（セキュリティ・新興技術センター、2022 年 3 月）、https://cset.georgetown.edu/publication/securing-ai-how-traditional-vulnerability-disclosure-mustadapt/。

[38] Tom Tseng、Euan McLean、Kellin Pelrine、Tony T. Wang、Adam Gleave、「Go AI は敵対的強靭性を備えることができるか？」、arXiv プレプリント arXiv:2406.12843 (2025)、https://arxiv.org/pdf/2406.12843。

[39] Lohn,「AI をハッキングする」

[40] ジェフ・ラーソン、スーリヤ・マットゥ、ローレン・カーチナー、ジュリア・アングウィン「COMPAS 再犯予測アルゴリズムの分析方法」ProPublica、2016 年 5 月 23 日、https://www.propublica.org/article/how-we-analyzedthe-compas-recidivism-algorithm; ロマン・ヤンポルスキー、「インシデント番号 11：ノースポイント・リスク・モデルズ」、人工知能インシデントデータベース、Responsible AI Collaborative、2016 年 5 月 23 日、https://incidentdatabase.ai/cite/11/。

[41] ラーソンら、「COMPAS の分析方法」; ヤンポルスキー、「インシデント番号 11」。

[42] ローマン・ルッツ、「インシデント番号 110：医療配分におけるアーカンソー州の不透明なアルゴリズムが受益者の診療時間を過剰に削減」、人工知能インシデントデータベース、Responsible AI Collaborative、2016 年 1 月 1 日、https://incidentdatabase.ai/cite/110。

[43] ルッツ、インシデント番号 110。

[44] コーア・ラム、「インシデント番号 503：Bing AI 検索ツールがユーザーへの脅威を宣言したと報じられる」、人工知能インシデントデータベース、Responsible AI Collaborative、2023 年 2 月 14 日、https://incidentdatabase.ai/cite/503/; ダニエル・アサートン、「インシデント番号 645：ジェミニのバイアスと社会技術的訓練の失敗が Google の評判を損なう傾向が見られる」、人工知能インシデントデータベース、Responsible AI Collaborative、2024 年 2 月 21 日、 https://incidentdatabase.ai/cite/645/;ローガン・B「インシデント番号 763：Grok AI チャットボットがトランプの入れ歯に関する根拠のない噂を拡散したと報じられる」人工知能インシデントデータベース、Responsible AI Collaborative、2024 年 8 月 13 日、https://incidentdatabase.ai/cite/763/。

[45] Roman Yampolskiy,「インシデント番号 47：LinkedIn 検索は男性名を優先する」, 人工知能インシデントデータベース, Responsible AI Collaborative, 2016 年 9 月 6 日, https://incidentdatabase.ai/cite/47/.

[46] 「インシデント番号 20：テスラ自動運転機能関連事故の事例集」人工知能インシデントデータベース、Responsible AI Collaborative、2016 年 6 月 30 日、https://incidentdatabase.ai/cite/20/。

[47] トーマス・ジャレラ、「インシデント番号 295：ニューヨーク市警の顔誤認識によるアップルストア窃盗容疑での不当な逮捕未遂」、人工知能インシデントデータベース、Responsible AI Collaborative、2018 年 11 月 8 日、https://incidentdatabase.ai/cite/295/; ダニエル・アサートン「事件番号 515：顔認識エラーによりジョージア州の男性が不当逮捕され、ルイジアナ州で 20 万ドルの和解金支払いに至った事例」人工知能インシデントデータベース、Responsible AI Collaborative、2022 年 11 月 25 日、

https://incidentdatabase.ai/cite/515/; Khoa Lam,「事件番号 517：アリバイがあるにもかかわらず、誤った顔照合により靴下窃盗で逮捕された男性」, 人工知能インシデントデータベース, Responsible AI Collaborative, 2018 年 2 月 15 日, https://incidentdatabase.ai/cite/517/; Khoa Lam,「インシデント番号 288：ニュージャージー州警察が顔認識技術で無実の黒人男性を誤逮捕」, 人工知能インシデントデータベース, Responsible AI Collaborative, 2019 年 1 月 30 日, https://incidentdatabase.ai/cite/288/; Daniel Atherton、「事件番号 630：メイシーズの顔認識エラーが誤逮捕とそれに続く刑務所内での性的暴行につながったとされるインシデント」、人工知能インシデントデータベース、Responsible AI Collaborative、2022 年 1 月 22 日、https://incidentdatabase.ai/cite/630/; ケイト・パーキンズ、「事件番号 439：デトロイト警察、誤った顔認識により黒人男性を不当逮捕」、人工知能インシデントデータベース、Responsible AI Collaborative、2019 年 7 月 31 日、https://incidentdatabase.ai/cite/439/。ダニエル・アサートン、「事件番号 598： ルイジアナ州警察の誤った顔認識技術によるジョージア州男性の誤った逮捕」、人工知能インシデントデータベース、Responsible AI Collaborative、2022 年 11 月 25 日、https://incidentdatabase.ai/cite/598/; ニッキー・デマコス、「インシデント番号 74：デトロイト警察が誤った顔認識技術により黒人男性を不当に逮捕」、 人工知能インシデントデータベース、 責任ある AI 共同体、2020 年 1 月 30 日、https://incidentdatabase.ai/cite/74/; コーア・ラム「インシデント番号 440：ルイジアナ州警察が誤った顔照合で黒人男性を不当逮捕」、人工知能インシデントデータベース、責任ある AI 共同体、2022 年 11 月 25 日、 https://incidentdatabase.ai/cite/440/; ダニエル・アサートン、「事件番号 692：ロンドン警視庁の顔認識技術がショーン・トンプソンを容疑者と誤認し逮捕に至ったと報じられる」、人工知能インシデントデータベース、Responsible AI Collaborative、2024 年 2 月 1 日、https://incidentdatabase.ai/cite/692; ダニエル・アサートン、「事件番号 816：ニューヨーク市警による管轄区域を超えた顔認識の誤認が、ニュージャージー州での不当逮捕と 4 年間の懲役刑につながった」、人工知能インシデントデータベース、Responsible AI Collaborative、2019 年 11 月 29 日、https://incidentdatabase.ai/cite/816。

[48] Anka Reuel、Amelia Hardy、Chandler Smith、Max Lamparth 他、「BetterBench：品質評価とベストプラクティス分析による情報に基づいたベンチマーク選択のための AI ベンチマークアセスメントのリポジトリ」、arXiv プレプリント arXiv:2411.12990 (2024)、 https://arxiv.org/abs/2411.12990。

[49] 「インシデント番号 162：ETS は、不正行為の規模を告発およびアセスメントするために、欠陥があるとされる音声認識証拠を使用したため、何千人もの人々が英国から国外追放された」人工知能インシデントデータベース、Responsible AI Collaborative、2014 年 1 月 1 日、http://www.incidentdatabase.ai/cite/162。

[50] エド・メインとリチャード・ワトソン、「数千人の人生を台無しにした英語試験」、*BBC*、2022 年 2 月 8 日、https://www.bbc.com/news/uk-60264106。

[51] チャルヴィ・ラストギ、ユンフェン・チャン、デニス・ウェイ他、「速い決断と遅い決断：AI 支援意思決定における認知バイアスの役割」、『*ACM* 人間とコンピュータの相互作用に関する論文集』第 6 巻第 83 号（2022 年 4 月）、1-22 頁、https://doi.org/10.1145/3512930。

[52] ショーン・マクレガー、「インシデント番号 172：ナルクスケアのリスクスコアモデルは妥当性確認不足で、バイアスリスクの高いデータで訓練されていたとされる」、人工知能インシデントデータベース、Responsible AI Collaborative、2020 年 7 月 1 日、https://incidentdatabase.ai/cite/172/。

[53] Maya Szalavitz,「痛みは耐え難かった。ではなぜ医師たちは彼女を拒んだのか？」 *Wired*, 2021 年 8 月 11 日, https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/.

[54] N.H. マックワース、「長時間の視覚探索における警戒心の崩壊」、『実験心理学季刊』第 1 巻第 1 号（1948 年 4 月）、6–21 頁、https://doi.org/10.1080/17470214808416738; Eric T. Greenlee, Patricia R.

DeLucia, and David C. Newton,「自動運転車両における運転者の警戒性：危険検知の失敗は時間の問題である」,『ヒューマンファクターズ』60 巻 4 号（2018 年 3 月）：465–476 頁, https://doi.org/10.1177/0018720818761711.

[55] 例えば参照：キャサリン・オルソン「インシデント番号 4：アリゾナ州で Uber 自動運転車が歩行者を死亡させた事故」、人工知能インシデントデータベース、Responsible AI Collaborative、2018 年 3 月 18 日、https://incidentdatabase.ai/cite/4; コア・ラム「インシデント番号 306：欧州高速道路でテスラの自動操縦 TACC がバンに衝突」、 人工知能インシデントデータベース、 Responsible AI Collaborative、2016 年 5 月 26 日、https://incidentdatabase.ai/cite/306 ;「インシデント番号 321：自動操縦中のテスラモデル X がカリフォルニア州高速道路の防護柵に衝突、運転手が死亡」、人工知能インシデントデータベース、Responsible AI Collaborative、2018 年 3 月 23 日、 https://incidentdatabase.ai/cite/321; Khoa Lam,「インシデント番号 353：フロリダ州で自動操縦中のテスラがトレーラートラックに衝突、運転手が死亡」人工知能インシデントデータベース、Responsible AI Collaborative、2019 年 3 月 1 日、https://incidentdatabase.ai/cite/353; ダニエル・アサートン「インシデント番号 434：自動運転モード中のテスラが急ブレーキ、トンネル内で多重衝突事故を引き起こす」人工知能インシデントデータベース、Responsible AI Collaborative、2022 年 11 月 24 日、https://incidentdatabase.ai/cite/434。

[56] 国家運輸安全委員会、「調査結果と勧告」、NTSB、2024 年 12 月 4 日、https://www.ntsb.gov/Advocacy/SafetyIssues/Pages/Vehicle-AutomationsInvestigative-Outcomes.aspx。

[57] クリスティアン・ヘッツナー、「米国がテスラの自動運転ソフトウェアのリコールに近づくにつれて、イーロン・マスクの規制上の苦境は深刻化する」『フォーチュン』2022 年 6 月 10 日 https://fortune.com/2022/06/10/elon-musk-tesla-nhtsainvestigation-traffic-safety-autonomous-fsd-fatal-probe/。

[58] ローレン・カーン、エメリア・S・プロバスコ、ロニー・キノシタ、「AI の安全性と自動化のバイアス」（セキュリティ・新興技術センター、2024 年 11 月）、https://cset.georgetown.edu/publication/ai-safetyand-automation-bias/。

[59] アストリッド・ベトラン、ラフィク・ベルーム、ジェームズ・R・イーガン、ウィンストン・マクスウェル「認知バイアス XAI 支援意思決定に与える影響：系統的レビュー」『*2022* 年 *AAAI/ACM* 人工知能・倫理・社会会議（*AIES '22*）論文集』（2022 年 7 月）：78-91 頁 https://doi.org/10.1145/3514094.3534164。

[60] Maja Jacobs、Melanie F. Pradier、Thomas H. McCoy Jr. 他、「機械学習による推奨が臨床医の治療選択に与える影響：抗うつ薬選択の事例」、『*Translational Psychiatry*』11 巻 108 号（2021 年 2 月）、https://www.nature.com/articles/s41398-02101224-x.

[61] 「インターネットユーザーのほとんどは、オンラインで健康情報を検索する際、検索エンジンから始める」（ピュー・リサーチ・センター、2006 年 10 月 29 日）、https://www.pewresearch.org/internet/2006/10/29/most-internetusers-start-at-a-search-engine-when-looking-for-health-information-online/。

[62] ショーン・マクレガー、「インシデント番号 139：監査により判明したアマゾンの検索・推薦アルゴリズムがワクチン誤情報を掲載した商品を優遇していた事実」、人工知能インシデントデータベース、Responsible AI Collaborative、2021 年 1 月 21 日、http://www.incidentdatabase.ai/cite/139/。

[63] プレルナ・ジュンジャとタヌシュリー・ミトラ、「アルゴリズムでキュレーションされたワクチン誤情報を巡る電子商取引プラットフォームの監査」、*2021* 年 *CHI* カンファレンス論文集（*CHI '21*）、第 186 号（2021 年 5 月 7 日）、1–27 頁、 https://doi.org/10.1145/3411764.3445250.

[64] キャサリン・オルソン、「事例 10：クロノス・スケジューリング・アルゴリズムがスターバックス従業員に経済的問題を引き起こしたとされる件」、人工知能インシデントデータベース、Responsible AI Collaborative、2014 年 8 月 14 日、https://incidentdatabase.ai/cite/10/。

[65] ジョディ・カントール「9 時から 5 時以外の働き」『ニューヨーク・タイムズ』2014 年 8 月 13 日、https://www.nytimes.com/interactive/2014/08/13/us/starbucks-workers-scheduling-hours.html。

[66] マイケル・トッパーとトシオ・フェラザレス、「警察技術の予期せぬ結果：ショットスポッターからの証拠」、ワーキングペーパー、2024 年 10 月 29 日、https://michaeltopper.netlify.app/research/jmp_michael_topper.pdf。

[67] ディロン・ライズマン、ジェイソン・シュルツ、ケイト・クロフォード、メレディス・ウィテカー「アルゴリズム影響評価：公共機関の説明責任のための実践的枠組み」（AI ナウ研究所、2018 年 4 月）、https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2。

[68] Simon Mylius と Jamie Bernardi、「AI による被害インシデントの追跡と分類」、MIT AI リスクリポジトリ、https://airisk.mit.edu/ai-incident-tracker.