ChatGPT

「Claude Mythos」関連文書流出（Anthropic）事案に関する分析報告書

エグゼクティブ・サマリー

本件は、AnthropicのWeb公開運用（外部CMS）に起因する「未公開コンテンツの意図せぬ公開可能状態」が発端となり、未発表モデル「Claude　Mythos」（同一モデルを指すとされる“Capybara”という新ティア名称も含む）の存在・一部仕様・公開方針・関連イベント情報などが、報道を通じて明るみに出た事案である。報道によれば、ブログ関連の未公開アセットが約3,000件、ログイン不要で到達できる形で置かれ、報道機関からの連絡後にAnthropicがアクセス制限した。 [1]

流出（＝公開状態にあった）素材は「モデルの重み」や「顧客データ」等ではなく、主にドラフト投稿・画像・PDF等の“公開準備物”であったとされ、Anthropic側も「外部CMSツールの設定における人的ミス」「公開候補の初期ドラフトであり、コア基盤やAIシステム、顧客データ、セキュリティ設計を含まない」旨を（報道を通じて）説明している。 [2]

一方で、ドラフトに記載された内容は「サイバー領域で他モデルを大きく先行し得る」「防御側の準備が追いつかない規模で脆弱性悪用が進む可能性」といったリスク認識を含み、段階的公開（まず防御側に早期アクセス付与）を示唆する。 [3]　これは近年の“高能力モデル×サイバー”の自己統制トレンド（例：高能力区分・段階的アクセス・監視強化）とも整合的で、他社も類似の枠組みを採用している。 [4]

市場面では、本件報道と同時期にサイバーセキュリティ関連株の下落が観測され、投資家が「AIによる攻防の非対称化（攻撃側優位）」を織り込み始めた可能性が示唆される（ただしマクロ環境等の交絡もあり、因果は限定的に解釈すべき）。 [5]

主要論点の確度（要約）

主張（要点）	確度	根拠（一次・準一次）
外部CMS設定ミスにより未公開アセット約3,000件が公開状態になった	高	当事者コメント（報道経由）＋独立研究者の件数確認 [6]
「Claude Mythos」および「Capybara」記載のドラフトが存在し、同一モデルを指す可能性が高い	高	ドラフト引用・説明（報道） [7]
記載内容（性能優位・“他AIよりサイバー能力で先行”など）は、社内ドラフト段階の主張であり最終仕様を保証しない	中	“ドラフト”である点、慎重公開方針の説明 [1]
「モデル重み・顧客データ等のコア資産の流出」ではない	中〜高	当事者コメント（報道経由） [2]
市場の短期反応（サイバー株下落）は本件に関連して語られた	中	市場記事（Investing.com等） [5]

事実関係と時系列

以下は、現時点で「一次（公式）／準一次（当事者発言を含む主要報道）」で裏取り可能な範囲に限定して整理した。流出ファイルの“原本URL”には意図的に触れない（拡散助長を避けるため）。代替として、主要報道が提示した限定的引用を根拠に要約する。 [8]

時系列テーブル（発端・流出経路・公開日時・関係者）

日時（現地表記→JST換算）	出来事（事実）	主な関係者	根拠	確度
2026-02-24（RSP v3.0 発効）	AnthropicがRSP v3.0を発効（リスクレポート等の枠組みを明記）	（公式文書）	[9]	高
2026-03-26（木）夕方まで（“Thursday evening”以前）	未公開ドラフトを含むデータストアが「公開検索可能」状態で存在していたと報道	—	[10]	中〜高
2026-03-26（木）	Fortune [11] がAnthropicに連絡→その後、公開検索・取得を制限したと報道	（当事者＋報道）	[8]	高
2026-03-26 22:25 ET（→03-27 11:25 JST）	未公開資産が約3,000件、CMS経由で公開状態だったとする“セキュリティ上の不備”記事公開	Beatrice Nolan [12] ほか	[13]	高
2026-03-26 22:27 ET（→03-27 11:27 JST）	「Claude Mythos」開発・早期顧客テスト中であることをAnthropicが認めた、とする続報公開	同上	[10]	高
2026-03-27 08:16（媒体表記、Investing.com）	サイバーセキュリティ株の下落が本件と関連づけて報じられる	Investing.com [14]	[15]	中
2026-03-27〜28（日本語圏）	日本語メディアが“CMS設定ミス・約3,000件・段階公開方針”として整理	すまほん!! [16]	[17]	中

流出経路フローチャート（推定を含む）

※以下は、報道にある事実（“外部CMS”“公開データストア”“公開がデフォルト”など）をベースに、一般化して模式化したもの。個別のAPI手順・探索方法など、再現に資する詳細は意図的に記述しない。 [2]

```
flowchart TD
A[社内: ブログ/広報/企画がドラフト・画像・PDFを外部CMSへアップロード] --> B[外部CMS/アセットストア]
B -->|公開がデフォルト/一部が未private| C[Publicに到達可能なデータキャッシュ/データストア]
C --> D[第三者(研究者/記者)が存在を認識・内容確認]
D --> E[報道機関が当事者へ連絡]
E --> F[Anthropicが検索/取得を制限(露出面の遮断)]
D --> G[報道: モデル存在・計画・リスク記述が公開]
```

流出文書の内容要約

公開された（と報道された）素材の範囲と性質

報道によれば、露出していたのはAnthropicサイトの発信（ブログ等）に紐づく未公開アセット群で、画像・ロゴ・バナー・PDF等を含む「約3,000件」規模だった。[6] また、当事者コメントとして「公開候補の初期ドラフト」「コア基盤やAIシステム、顧客データ、セキュリティ設計を含まない」という位置づけが示されている。[2]

内容要約テーブル（技術仕様・能力・訓練データ・内部方針・商業計画）

カテゴリ	要旨（報道ベース）	重要抜粋（短引用）	含意（リスク/機会）	確度
モデル名称・ティア構造	「Claude Mythos」と「Capybara」が同一モデルを指す可能性。CapybaraはOpusより上位の新ティアとして説明	“‘Capybara’ is… larger and more intelligent than our Opus models”[10]	製品ライン再編・価格帯再設定・提供形態（限定公開）の可能性	高
能力（推論/コーディング/サイバー）	Opus 4.6より、コーディング・学術推論・サイバー関連タスクなどで高スコアと記載	“dramatically higher scores… coding… reasoning… cybersecurity”[10]	攻撃・防御双方の自動化が進む可能性（デュアルユース加速）	中〜高
“最強”表現	社内ドラフト上で“最も強力”と形容	“by far the most powerful AI model we’ve ever developed”[18]	マーケ/広報文脈の誇張可能性も。だが当事者が“step change”と表現	中
安全性/セキュリティ懸念（サイバー）	「他AIよりサイバー能力で先行」「防御側を上回る速度で脆弱性悪用の波」等を警告	“far ahead of any other AI model in cyber capabilities”[18]	攻撃スケール化・ゼロデイ探索効率化・“低熟練者の底上げ”懸念	中
公開戦略（段階的アクセス）	まず防御側組織に早期アクセスを与え、耐性向上の“先行時間”を確保する方針	“releasing it in early access to organizations… head start”[7]	“Trusted access”型ガバナンスの採用を示唆（規制との親和性）	中〜高
運用コスト	計算資源負担が大きく、一般提供が未定/高コストと記載	“very compute intensive… very expensive”[19]	価格形成・提供形態（APIのみ/上位プランのみ等）に影響	中
商業計画（エンタープライズ）	欧州大企業CEO向け招待制イベント（英国内マナー/ホテル等）資料が含まれる	「招待制CEOサミット」「未公開Claude能力の体験」等[7]	大企業販売の加速、規制/政策対話の場として利用可能	中

カテゴリ	要旨（報道ベース）	重要抜粋（短引用）	含意（リスク/機会）	確度
訓練データ	流出素材から訓練データ詳細が出たとは報道されていない（少なくとも主要記事には明示なし）	—	本件から“学習データの具体”を断定するのは不可	低
内部方針（安全枠組み）	（流出とは別に）AnthropicはRSP v3.0でRisk Report等の透明性枠組みを宣言	Risk Reportsを「オンライン公表（部分的redaction）」等 [20]	今後、サイバー能力の扱いを“公式文書”で裏付ける圧力が高まる	高

訓練データに関する「背景情報」と限界（重要な不確実性）

本件の“流出ドラフト”それ自体が訓練データの詳細を含むかは、主要報道からは確認できないため、本報告書では断定しない。 [1]
ただし、Anthropicは既存モデルのシステムカードで「公開Web情報＋第三者提供の非公開データ＋ラベリング/委託データ＋オプトインユーザーデータ＋社内生成データ」等の組成を明示しているため、将来モデルでも“同種のカテゴリ構成”が継続する可能性はある（推測であり、Mythosに直接適用は不可）。 [21]

技術的評価

記載内容の信頼性評価（ドラフト＝確定仕様ではない）

本件の“能力・リスク”記述は、報道によれば「公開前のドラフト投稿（構造化データ、見出し、公開日を含む）」であり、製品版の仕様やベンチマーク詳細が確定していることを意味しない。 [8]
一方で、「モデルが開発・早期顧客テスト中で、推論・コーディング・サイバーで意味のある前進がある」という点は、当事者コメントとして報道されており、事実関係としての信頼性は相対的に高い。 [22]

評価としては、

- **存在・開発段階（step change、早期アクセス顧客あり）**：高信頼
- **“他AIよりサイバー能力で先行”など相対比較の強い文言**：中信頼（社内評価/マーケ文脈の可能性、検証指標不明）
- **“防御側を上回る速度で脆弱性悪用の波”という将来予測**：中〜低信頼（前提条件・攻撃者制約が大きい）

  が妥当である。これは、NISTがデュアルユース基盤モデルの誤用リスクを論じる際に「能力測定が現実被害に直結しない」「ドメイン間で能力が移りにくい」などの不確実性を強調している点とも整合する。 [23]

技術的インパクト（サイバー領域の“攻防スピード”）

ドラフトが示唆するリスクの核心は「脆弱性発見〜悪用の自動化が、攻撃者側のスループットを押し上げ、防御側のパッチ/検知の速度を上回り得る」という点にある。 [18]
この“攻防スピード差”は、（モデル単体の賢さだけでなく）ツール実行・長時間タスク・システム統合によって増幅される。実際、他社もサイバー能力が上がったモデルに対して「高能力区分」「監視」「限定アクセス」等を組み合わせる姿勢を示している。 [4]

また、防御側の観点では、MITREがAI組込みシステムの攻撃面拡大（データ依存・学習過程）やサプライチェーン不透明性を論じ、AI特有の脅威を系統的に洗い出す枠組み（NIST・ATLAS等との接続）を提示して

いる。[24]
Mythos級の能力が現れるなら、**“AIを守る”**（モデル/データ/推論パイプラインの保護）と**“AIで守る”**（自動監査、継続的脆弱性探索、攻撃シミュレーション）の両輪が、従来以上に不可欠になる。

セキュリティ/安全性上の懸念（本件“漏えい”そのもの）

本件は「外部CMSのアセット公開設定がデフォルトでpublicになり得る」「ユーザがprivate設定を忘れると未公開物が露出する」という運用設計・ガバナンス課題を示した。[2]
研究開発企業にとって、CMSは“広報ツール”であると同時に、未発表計画・契約前提資料・人事/イベント資料が集積する**準レポジトリ**になりやすい。今回、タイトル上で「社員の育休（parental　leave）」を示すアセットが含まれたとされる点は、情報分類とアクセス制御が“モデルの重み”だけでなく周辺業務にも及ぶべきことを示唆する。[2]

法的・倫理的論点

機密情報保護（トレードシークレット、契約、注意義務）

米国法の一般論として、営業秘密の不正取得に対し民事救済を認める枠組みが存在する（例：Defend　Trade Secrets Actに基づく民事請求）。[25]
ただし本件は「第三者が侵入した」というより「当事者側の設定不備でpublic到達可能になった」形であり、**“不正取得（misappropriation）”の立証構造**や、公開状態が営業秘密性に与える影響は、個別事情（アクセス制限の有無、利用規約、閲覧者の取得態様、公開期間、拡散度合い等）に強く依存する。[2]
倫理的には、報道価値があるとしても「未公開・限定公開の資料」を二次拡散する行為は、被害を増幅しうるため、責任ある取り扱い（最小限引用、個人情報の秘匿、再現可能な取得手順の非開示）が要請される。

労働法・個人情報保護（HR情報が含まれた可能性）

報道では、内部向け画像に「社員の parental leave」を示すタイトルが含まれた可能性が示されている。[2]
これが特定個人に結びつく情報（氏名、ID、健康情報等）を含む場合、法域によっては個人情報保護法制上のインシデント対応（評価・通知）が必要となり得る。
EU域内に関係する個人データが含まれ得る場合、GDPRは一定条件下で監督機関への通知（一般に72時間以内）などを定める。[26]
現時点では、当該アセットが個人データに該当するか、どの法域が適用されるか（英国/欧州サミット資料、米国拠点等）は公開情報だけでは確定できないため、企業側は法務・プライバシー部門による**“含まれていた情報の分類”**を最優先で行うべきである。[27]

規制対応（EU AI ActのGPAI義務と“システミックリスク”）

EU　AI　Actは適用時期・義務が段階化されており、欧州委員会の説明では「GPAI（汎用AI）モデル提供者の義務は2025-08-02から適用」「全面適用は2026-08-02」などが明示されている。[28]
Mythosが実際にどの市場で提供され、GPAIとしてどの区分（特に“systemic　risk”相当）に該当し得るかは未確定だが、少なくとも“高能力×サイバー”が制度上の注視対象になる方向性は強い。これはNISTがデュアルユース基盤モデルの誤用（攻撃的サイバー等）をリスク領域として明示している点とも概念的に合致する。[29]

反応・類似事例比較・影響分析

関係者の反応（時系列）

時点	主体	反応（要旨）	根拠	備考/確度
2026-03-26	Anthropic（広報）	外部CMSの人的ミスでドラフトが到達可能になった／モデル開発は事実、慎重にリリース	“human error”“being deliberate”など [30]	高
2026-03-26	研究者（Roy Paz [31]、Alexandre Pauwels [32]）	未公開アセット約3,000件を確認（報道機関の依頼で検証）	[7]	中〜高
2026-03-27	市場アナリスト（Adam Tindle [33] 等）	AIで未知の攻撃が高速化し、従来型防御（シグネチャ等）が圧迫されうる	[34]	中
同上	市場アナリスト（Kirk Materne [35]、Adam Borg [36]）	“長い調整”“究極のハッキングツール化”など懸念、ただし防御強化需要も示唆	[37]	中
2026-03-28	日本語メディア	“約3,000件”“人的ミス”“防御側へ早期アクセス”として整理	[17]	中

市場への短期影響（観測事実と解釈の分離）

Investing.comは、サイバーセキュリティ株が下落し、個別銘柄ではCrowdStrike [38]、Palo Alto Networks [39]、Zscaler [40]、Okta [41]、SentinelOne [42]、Fortinet [43] 等の下落が列挙されている。 [15]
同時に日本語記事でも「4〜7%程度の下落」や暗号資産市場のリスクオフ文脈が語られている。 [44]
ただし、株価変動は多要因であるため、本件が寄与したとしても**“単独因果”を断定しない**のが適切である（投資行動の確定的判断材料にはならない）。

類似事例比較（AI関連の“漏えい/露出”の型）

事例	漏えい対象	経路（型）	影響の特徴	主な根拠
Meta [45] のLLaMA流出（2023）	モデル重み（ウェイト）	配布制限下での不正再配布（torrent等）	「一度出た重みは回収困難」→派生モデル・悪用/研究促進の両論	[46]
OpenAI [47] のChatGPT障害/情報露出（2023）	他ユーザーのチャットタイトル等	キャッシュ/OSSバグ起因の露出	サービス停止・公式説明・影響範囲の特定が中心	[48]
Samsung Electronics [49] の社内情報流出（2023）	社内コード/会議情報	従業員が外部LLMに入力（運用・教育問題）	企業側の利用禁止/ガバナンス強化の契機	[50]

本件（Mythos）は上の分類で言えば、「モデル重み流出」ではなく「周辺運用（CMS/公開準備）からの露出」であり、

- **技術的被害の深刻度（重み流出ほど不可逆ではない）**と、
- **戦略/計画の露出（競争・社会的影響の大きさ）**
  が同時に起きうる点が特徴である。 [6]

短中長期の業界影響（リスクと機会）

短期的には、Anthropicの「安全重視」ブランドに対するレピュテーション圧力と、エンタープライズ顧客（早期アクセス顧客を含む）の信頼・契約交渉への影響が焦点となる。 [7]

中期的には、“高能力×サイバー”モデルに関して、段階的アクセス（Trusted access）、監視強化、研究者との協働（防御側先行）といった**ガバナンス実装が標準化**する可能性がある。実際に、OpenAIはサイバー領域で「High capability」区分、強化セーフガード、限定アクセス（Trusted Access for Cyber）などを発表している。 [4]

長期的には、EU AI ActのGPAI義務適用・執行開始（2025〜2026）と重なる形で、各社のリスク報告・外部評価・インシデント対応が、より“規制適合”として外部から監査される局面に入る。 [51]
また、NISTやMITREが提示するように、AIは新しい攻撃面（データ依存、供給網、モデルの不透明性）を持つため、セキュリティ産業は「AIが脅威を増やす」という側面だけでなく、「AIを前提とした防御設計・評価」の市場（AIネイティブSOC、継続的検証、AIサプライチェーン評価等）を拡大させうる。 [52]

関係者と影響のER図（概念）

```
erDiagram
  COMPANY ||--o{ SYSTEM : "uses"
  SYSTEM ||--o{ ASSET : "stores"
  ASSET }o--|| PUBLIC_STORE : "exposed_via"
  MEDIA ||--o{ PUBLIC_STORE : "accessed/verified"
  RESEARCHER ||--o{ PUBLIC_STORE : "assessed"
  MEDIA ||--o{ COMPANY : "notified"
  COMPANY ||--o{ PUBLIC_STORE : "restricted"
  COMPANY ||--o{ EARLY_ACCESS_CUSTOMER : "tests_with"
  COMPANY ||--o{ CYBER_DEFENDER : "plans_early_access"
  MODEL ||--o{ CYBER_RISK : "enables"
  REGULATOR ||--o{ COMPANY : "oversight/obligations"
```

推奨される対応策

企業（Anthropicを含む開発企業・クラウド/ベンダ）向け

第一に、外部CMSやアセットストアは「広報」ではなく「機密情報の一時保管庫」になり得る前提で、**デフォルト非公開（private by default）**・自動分類（機密/公開）・公開前ゲート（承認ワークフロー）を必須化する。今回のように“publicがデフォルト”であれば、設計として事故が起きやすい。 [2]

第二に、RSP v3.0やRisk Reportが志向するような「評価・緩和策・公開の整合性」を、モデル本体だけでなく周辺運用（広報、イベント、採用、取引）にも拡張し、**“情報のサプライチェーン”**を統制する。 [53]

第三に、“高能力×サイバー”モデルについては、NISTが論じるデュアルユース誤用（攻撃的サイバー等）を前提に、段階公開、監視、外部評価、透明性（ただし悪用助長情報は最小化）の設計を標準機能として組み込む。 54

第四に、防御側の実装指針としては、MITREのようなAI特有の脅威整理（データ汚染、供給網、機密情報の埋め込み等）を踏まえ、モデル・データ・CI/CD・運用監視まで含めた統合的コントロール選択が重要になる。 24

規制当局向け（EU/米国等）

GPAI提供者義務が既に適用段階に入っているEUでは、実務ガイダンスに沿って「システミックリスクモデルの届出/協働」「透明性」「インシデント時の説明責任」を、実装面（監査可能な証跡）で求めることが最も効果的である。 55

また、デュアルユース・誤用リスクについては、NISTが示すように技術だけでなく社会的要因（攻撃者資源、防御側の成熟度）も含めた評価が必要であり、単発のベンチマークで規制閾値を定めることには慎重さが求められる。 56

研究者・メディア向け（責任ある取り扱い）

「脆弱性の再現手順」や「未公開資料の所在」を拡散させない一方で、社会的意義のある論点（サイバー能力、段階公開、評価枠組み）のみを検証可能な形で議論することが望ましい。これは本件報道が当事者コメント・限定引用に留めている点とも整合する。 8

最終エグゼクティブ推奨（短い行動案）

- **外部CMS/アセットストアを“機密情報システム”として再分類**し、private-by-default＋公開前承認（強制）へ移行する。 2
- **“高能力×サイバー”モデルはTrusted access＋監視＋外部評価を前提**に段階公開し、誤用リスク管理の証跡を残す。 57
- **RSP/Risk Report等の透明性枠組みを、モデル以外（広報/イベント/HR）にも拡張**し、情報サプライチェーン全体を統制する。 53
- **規制側はEU AI ActのGPAI義務運用を“監査可能な実装要件”として具体化**し、事後対応ではなく事前の統制設計を促す。 55
- **研究者・メディアは責任ある開示（最小限引用・再現手順非公開）を徹底**し、社会的に重要な論点に限って検証可能性を高める。 8

参考資料一覧（原典リンク優先、日英混在）

主要報道（準一次）

- Fortune 11 （2026-03-26）：未公開アセット約3,000件の露出、CMS設定、当事者コメントなど。 2
- Quartz 58 （2026-03-26/27相当）：研究者検証、イベント情報、段階公開の要旨。 59
- Investing.com 14 （2026-03-27）：市場反応、アナリストコメント。 60

公式・一次（ガバナンス/リスク管理）

- Anthropic：Responsible Scaling Policy v3.0（2026-02-24発効、PDF）。 61
- Anthropic：Risk Report（Feb 2026、PDF）。 62

- Anthropic：Claude 4 System Card（訓練データのカテゴリ等、PDF）。 21
- OpenAI：GPT-5.3-Codex紹介（サイバー高能力区分、Trusted Access for Cyber等）。 63

公的ガイダンス・規制（一次）

- European Commission 64 ：EU AI Actの適用タイムライン（GPAI義務時期等）。 28
- National Institute of Standards and Technology 65 ：Dual-use foundation modelの誤用リスク管理ガイド（NIST AI 800-1、IPD）。 56
- GDPR（EUR-Lex）：データ侵害通知（Article 33等）。 26
- 米国（US Code）：Defend Trade Secrets Act関連条文（18 U.S.C. § 1836等）。 25

防御側フレームワーク（一次）

- MITRE 66 ：SAFE-AI（AI組込みシステムのセキュリティ枠組み、ATLAS連携）。 24

類似事例（比較用）

- The Verge 67 ：LLaMA流出報道（2023）。 68
- OpenAI：ChatGPT障害/露出の公式説明（2023-03）。 69
- Reuters 70 ：ChatGPTの露出報道（2023-03）。 71
- Forbes 72 ：Samsungの社内情報流出と対策（2023）。 73
- The Japan Times 74 ：Samsungの生成AI利用制限報道（2023）。 75

日本語圏の整理（補助）

- すまほん!! 16 ：CMS設定ミス・約3,000件・段階公開方針などの日本語要約。 76
- Yahoo!ファイナンス（CoinPost配信）：市場警戒と日本語要約（ただし“開発用ページ”など二次情報が混在）。 44

---

1 3 7 8 10 18 22 27 30 32 33 40 47 49 57 64 65 66 67 https://fortune.com/2026/03/26/anthropic-says-testing-mythos-powerful-new-ai-model-after-data-leak-reveals-its-existence-step-change-in-capabilities/
https://fortune.com/2026/03/26/anthropic-says-testing-mythos-powerful-new-ai-model-after-data-leak-reveals-its-existence-step-change-in-capabilities/

2 6 13 14 36 38 39 https://fortune.com/2026/03/26/anthropic-leaked-unreleased-model-exclusive-event-security-issues-cybersecurity-unsecured-data-store/
https://fortune.com/2026/03/26/anthropic-leaked-unreleased-model-exclusive-event-security-issues-cybersecurity-unsecured-data-store/

4 63 https://openai.com/index/introducing-gpt-5-3-codex/
https://openai.com/index/introducing-gpt-5-3-codex/

5 11 15 19 31 34 37 42 43 60 https://www.investing.com/news/stock-market-news/cybersecurity-stocks-plunge-as-anthropics-claude-mythos-leak-sparks-ai-fear-4584897
https://www.investing.com/news/stock-market-news/cybersecurity-stocks-plunge-as-anthropics-claude-mythos-leak-sparks-ai-fear-4584897

9 12 20 53 61 74 https://anthropic.com/responsible-scaling-policy/rsp-v3-0
https://anthropic.com/responsible-scaling-policy/rsp-v3-0

16 17 76 https://smhn.info/202603-anthropic-claude-mythos-leaked-cms-misconfiguration
https://smhn.info/202603-anthropic-claude-mythos-leaked-cms-misconfiguration

21 https://www.anthropic.com/claude-4-system-card
https://www.anthropic.com/claude-4-system-card

23 29 52 54 56 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf

24 https://atlas.mitre.org/pdf-files/SAFEAI_Full_Report.pdf
https://atlas.mitre.org/pdf-files/SAFEAI_Full_Report.pdf

25 https://www.law.cornell.edu/uscode/text/18/1836
https://www.law.cornell.edu/uscode/text/18/1836

26 35 58 72 https://eur-lex.europa.eu/legal-content/EN-SV-ES/TXT/?uri=CELEX%3A32016R0679
https://eur-lex.europa.eu/legal-content/EN-SV-ES/TXT/?uri=CELEX%3A32016R0679

28 51 https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

41 55 https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers
https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers

44 https://finance.yahoo.co.jp/news/detail/d9d6ea18aaf4a1f0b8a5ee41d3495bbddb2e984a
https://finance.yahoo.co.jp/news/detail/d9d6ea18aaf4a1f0b8a5ee41d3495bbddb2e984a

45 75 https://www.japantimes.co.jp/news/2023/05/02/business/tech/samsung-bans-chatgpt-workplace-use/
https://www.japantimes.co.jp/news/2023/05/02/business/tech/samsung-bans-chatgpt-workplace-use/

46 68 https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse
https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse

48 69 https://openai.com/index/march-20-chatgpt-outage/
https://openai.com/index/march-20-chatgpt-outage/

50 73 https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/
https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/

59 https://qz.com/anthropic-claude-mythos-data-leak
https://qz.com/anthropic-claude-mythos-data-leak

62 https://anthropic.com/feb-2026-risk-report
https://anthropic.com/feb-2026-risk-report

70 71 https://www.reuters.com/technology/chatgpt-owner-openai-fixes-significant-issue-exposing-user-chat-titles-2023-03-22/
https://www.reuters.com/technology/chatgpt-owner-openai-fixes-significant-issue-exposing-user-chat-titles-2023-03-22/