本稿は J-AISI の公開済み資料「AI システムに対する既知の攻撃と影響」[40]の詳細版である“Securing AI Systems: A Guide to Known Attacks and Impacts" (arXiv:2506.23296v1 [cs.CR]) の参考日本語訳である。

AI システムを守る

既知の攻撃と影響についての手引き

Securing AI Systems: A Guide to Known Attacks and Impacts (Japanese Version)

桐淵 直人[1,2]　銭谷 謙吾[1,2]　瀬光 孝之[1,2]

NAOTO KIRIBUCHI,　KENGO ZENITANI,　TAKAYUKI SEMITSU

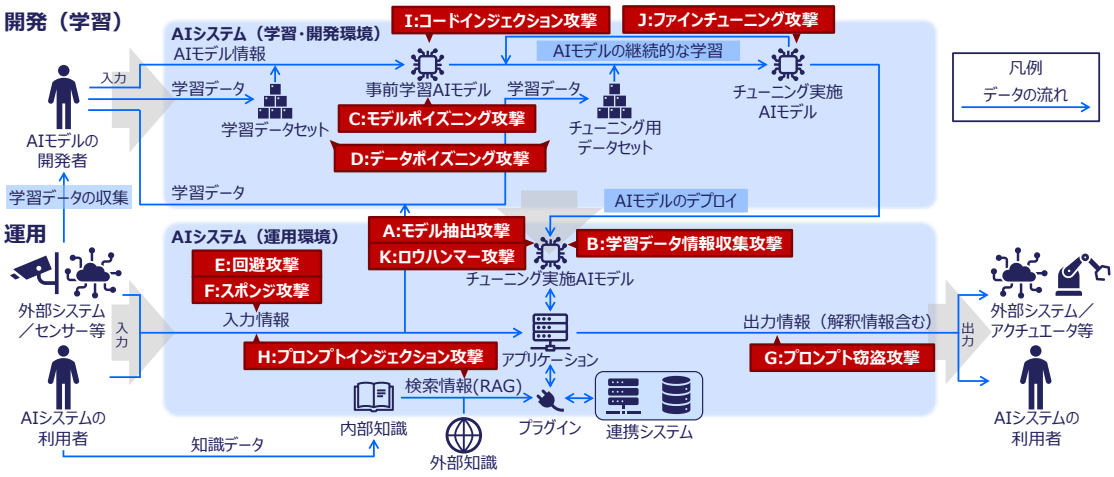

図 1　AI システムに対する既知の 11 種の攻撃（A~K）の概要

情報システムに組み込まれた人工知能（AI）は、AI 特有の脆弱性を悪用するセキュリティ上の脅威に直面している。本稿は、予測 AI システムおよび生成 AI システムに特有の攻撃について、分かりやすい全体像を提供するものである。我々は 11 の主要な攻撃タイプを特定した上で、個々の攻撃手法を、情報漏洩、システム侵害、計算資源浪費といった影響に明示的に関連付ける―これらの影響は、機密性、完全性、可用性（CIA）のセキュリティ 3 要件に対応する。本稿の目的は、AI セキュリティの専門知識を持たない者も含め、研究者・開発者・セキュリティ実務者・政策立案者が、AI 特有のリスクを認識し、効果的な対策を実施するための基礎知識を身につけられるよう支援し、ひいては、AI システムの全体的なセキュリティを強化することにある。

参考キーワード: AI セキュリティ・敵対的機械学習・予測 AI・生成 AI

---

[1] AI セーフティ・インスティテュート（Japan AI Safety Institute, J-AISI）

[2] 情報処理推進機構（Information-technology Promotion Agency, IPA）

1. 導入

人工知能（AI）の分野は近年急速な進歩を遂げ、医療、金融、コンテンツ生成を含む多様な領域で革命的な変化をもたらしている。技術革新は AI の潜在的な応用範囲を拡大しただけでなく、その社会的影響力も著しく増大させた。AI システムが重要なインフラや意思決定プロセスに組み込まれるにつれて、総称して AI セーフティと呼ばれる、安全で信頼でき、倫理的な運用を確保することへの世界的な認知と要求が著しく高まっている。

世界的な懸念のこうした高まりに応え、日本は 2023 年に採択された包括的な政策枠組みである広島 AI プロセス[66]を主導することで先駆的な役割を担ってきた。時を同じくして、英国と米国は AI 関連のリスクに対処するため、国家の取り組みとして、AI セーフティ・インスティテュート（AI Safety Institute）を設立した[1,2]。これらの世界的な動向に呼応し、日本においても 2024 年 2 月に AI セーフティ・インスティテュート（J-AISI）が設立された。J-AISI は、AI 技術の責任ある社会実装に向けた国内外の協力を促進し、技術的知見を集約するという使命のもと、日本の AI セーフティに関する中核的ハブとして機能する。J-AISI の事務局は情報処理推進機構（IPA）内に設置[39]され、これにより J-AISI の活動に求められる調整と支援を円滑なものにしている。

広範な AI セーフティの領域において、AI システムの機密性・完全性・可用性を損なうよう仕組まれた悪意のある活動から AI システムを保護することに焦点を当てた AI セキュリティが極めて重要な分野として台頭している。AI システムがますます機密性の高いデータを取り扱い、重要な社会的機能を担うようになるにつれ、AI セキュリティの不備がもたらす影響も拡大する。具体的には、プライバシー侵害、運用停止、経済的損害、公共の安全への脅威といった深刻な結果につながる可能性がある。したがって、AI セキュリティのリスクとそれを緩和するための実践的な戦略を明確に理解することは、AI の専門家だけでなく開発者・セキュリティ実務者、そして多様なセクターで AI システムの導入・運用・管理に責任を負うステークホルダーにとって今や不可欠である。

本稿は、AI セキュリティの専門知識を持たない者を含む幅広いステークホルダー、中でも特にセキュリティ実務者に対し、図 1 に示すような AI システムへの攻撃の包括的で極力分

---

[1] 英 AI Safety Institute は 2025 年 2 月 14 日に AI Security Institute に改名した。[45]

[2] また、米 AI Safety Institute は Center for AI Standards and Innovation (CAISI)に改名されることが 2025 年 6 月 3 日に公表された。[63]

かりやすい概要を提供することで、AI セキュリティに貢献する。というのも、AI セキュリティには膨大な研究があり、この分野に不慣れな実務者が、その概況や背景となる文脈について、一貫性のある包括的な理解を深めるのが必ずしも容易ではないからである。AI セキュリティの包括的な理解を支援するためには、実用的なガイドが必要である。なお、最新かつ最先端の攻撃に関する包括的なサーベイは、本稿のスコープを超えるものとして扱わない。

本稿の際立った特徴は、第一に AI システムへの攻撃が及ぼすセキュリティ上の影響に焦点を当てていることと、第二に、セキュリティ上の影響と攻撃タイプの間、並びに、異なる攻撃タイプの間の関係性を示す、直感的・視覚的な要約に重きを置いていることである。本稿は、先行研究により確立された各攻撃タイプとそれぞれが AI システムに及ぼす具体的な影響との結び付きを明示することで、セキュリティマネジメントの実務に有用な洞察を読者に提供する。開発者や実務者は、構造化されたこの知見を活用することで、リスクアセスメントやレッドチーム演習を実施し、更に、自身の AI システムに関わりの深い重大な攻撃ベクトルを特定できる。

本稿では、参照した攻撃文献を、モダリティ（例：テキスト、画像、表形式）、モデルタイプ（例：ニューラルネットワーク、ロジスティック回帰、決定木）、攻撃者の能力（攻撃を成功させるのに必要な前提条件）という切り口で整理して、表 4 にまとめている。この表は、広範なサイバーセキュリティの文脈の中で、読者が現実的なリスクシナリオを評価する上で実用的な指針を提供する。

まとめると、本稿の主な貢献は以下の通りである：

- AI 関連の多様な関係者による世界的に確立されたリファレンス類を参照することで、AI システムに対する 11 の攻撃タイプを特定している。
- 各攻撃タイプと AI システムへの個別の影響の関連を明示することで、レッドチーム演習を含む実践的なリスクアセスメントを支援する。
- 各攻撃の実践に際し攻撃者が必要とする能力（攻撃の前提条件）を特定し、AI セキュリティのシナリオとリスクの現実的な評価を支援する。
- 攻撃のメカニズムとシステムに及ぼすセキュリティ上の影響についての、AI セキュリティの専門知識を持たない開発者やセキュリティ実務者に向けた、分かりやすく直感的な要約を提供する。
- 攻撃に関する参考文献をデータモダリティと AI モデルタイプに従って整理し、更なる詳細の学習を読者が進める際の簡潔なロードマップを提供する。

以後の本稿の構成は次の通りである。第 2 節では、AI セキュリティの関連研究を概観し、既存のフレームワークや文献との対比で我々の貢献の位置づけを明確にする。第 3 節では、本稿で想定する AI システムの一般的なアーキテクチャを詳述し、本稿で検討する攻撃のスコープを定義する。第 4 節では、CIA セキュリティ 3 要件に対応付けながら、11 の主要な攻撃タイプの分類を導入し、それらが AI システムに及ぼす潜在的な影響を論じる。第 5 節では、特定された各攻撃タイプについて、文献による裏付けを示しつつ、そのメカニズム、攻撃者が必要とする能力（攻撃成立の前提条件）、影響を受ける AI モダリティとモデルタイプを詳細に説明する。最後に、第 6 節において我々の主要な知見を要約し、将来の研究の方向性を示唆することで、本稿を締めくくる。

2. 関連研究

この節では、AI セキュリティの文脈における本稿の位置づけを明らかにするべく、AI システムに対する敵対的脅威についての関連研究を概観し、それらの貢献と限界を指摘する。この節で参照する既存のフレームワークと文献のなした多大な貢献は、AI セキュリティに関する世界的な議論を先導した礎であって、我々はこれを深く尊重する。本稿の目的は、これらの研究を否定したり置き換えたりすることではなく、追加の洞察を提供することで、先行する成果の実応用に貢献することにある。先行する成果の間に見られる相違点と補足的な視点とを明らかにすることで、我々はこれらの研究全体に対する理解を深める一助を提供すると共に、特にセキュリティ実務者の観点から主要な攻撃タイプの特徴を指摘し、国際的な議論の将来的な発展に資することを目指す。

2.1. 学術的サーベイ

予測 AI システムに関連する敵対的脅威と防御手法については、いくつか学術的サーベイが体系的に分析している。

Kong らは、モダリティ特化の攻撃、具体的には画像、テキスト、マルウェアのドメインを対象とした攻撃に焦点を当てた包括的な分類を提供し、典型的な攻撃アルゴリズムとそれに対する防御策を詳述している[44]。Oseni らは敵対的攻撃を分析し、機械学習の概要と攻撃の数学的側面を説明している[69]。加えて、Hu らは AI システムのライフサイクルに基づいて敵対的攻撃を分類することで、脅威を様々な開発段階（例：学習データの収集、前処理、モデルの学習、推論）に対応付け、ライフサイクル特有の脆弱性と防御策を明らかにしている[37]。Chen と Babar もライフサイクルの視点を採用しているが、ソフトウェア工学の実践とセキュアな開発方法論にまで分析を拡張しており、堅牢な AI システムの構築に関する

ガイダンスを提供[17]している。

これらのサーベイは包括的ではあるものの、主に予測 AI システムを取り上げており、生成 AI システムに関する洞察を欠いている。これに対し最近のサーベイは、生成 AI の主要なカテゴリである大規模言語モデル（LLM）に焦点を当てることで、このギャップに対処し始めている。

Yao らは、LLM の脆弱性を AI 固有のものと AI 固有でないものとに分類し、モデル抽出やリモートコード実行などのリスクを特定している[98]。Shayegani らはマルチモーダル攻撃を含む LLM に対する敵対的攻撃を体系的に調査し[77]、その一方で Dong らは、AI との会話の安全性について、攻撃の方法論と対応する防御策を含む形で詳細に検討している[25]。Yan らは LLM のプライバシーリスクと対策について詳細な分析を提供し[97]、Das らはジェイルブレイキングやデータポイズニングといった更なる様々なセキュリティ脅威を明らかにしている[22]。対照的に、Sun らは最近、敵対的生成ネットワーク（GAN）と変分オートエンコーダ（VAE）を調査し、主にモデル中心の脆弱性に焦点を当てている[85]。ただ、彼らの分析はモデル中心の脆弱性に対処するものであり、より広範なシステム統合の問題はほとんど考慮されていない。

上記のように、既存のサーベイは予測モデルか生成モデルのいずれかに焦点を当てつつもモデルレベルの議論に留まっており、生成 AI システムが有するシステム的な脆弱性を含む包括的研究は依然として必要である。また、これらの学術的サーベイは AI 研究者を対象としていて実務者を対象としておらず、その知見を NIST AI 100-2e2025 や MITRE ATLAS™のような実務的フレームワークとの明示的な対比は論じていない。

以下の項では、主要なフレームワークを実務者の視点から概観し、上記で議論した学術的サーベイとの違いおよび補完的な役割を浮き彫りにする 。

2.2. NIST AI 100-2e2025

近年の注目すべき成果として、米国国立標準技術研究所（NIST）による「敵対的機械学習：攻撃と緩和策の分類と用語」（NIST AI 100-2e2025）[90]と題された改訂版の報告書がある。以前の版[89]と比較して、この最新版では文献レビューの範囲が参考文献 400 件以上に拡大しており、初期の概念から最新の研究に至るまでの歴史的変遷を徹底的に解説している。この報告書はまた、AI セキュリティ用語の定義を提供する用語集も兼ねている。本稿でもこの NIST の報告書を参照しており、NIST の報告書で引用されている多数の文献は、我々の分析においても重要な役割を果たしている。

他方で、新たに導入された索引によって旧版より参照しやすくなったものの、NIST の報告書が提示する分類体系は依然として複雑である。NIST の報告書では攻撃を、まずは対象となる AI システムのタイプ（すなわち、予測 AI システムと生成 AI システム）によって、次に攻撃者の目的（すなわち、可用性、完全性、プライバシー、悪用、サプライチェーン攻撃）によって分類している。しかし、多くの攻撃手法はこれらの分類を横断しているため、報告書内での同様の記述の繰り返しを招いている。例えば、間接プロンプトインジェクション（NISTAML.015）と出力ミスアラインメント（NISTAML.027）は技術的メカニズムとしては同一の攻撃であるが、目的の観点から複数の異なる説明が与えられている。同様に、これも概念としては同一のモデルポイズニング攻撃が、対象となる AI システムのタイプや目的に応じて、異なる ID（NISTAML.011,026,051）で分類されている。記載のこうした重複は、特に実務者にとって、同分類体系の実務上の有用性を損ねる可能性がある。

対照的に、我々は簡素化した分類体系を意図的に採用している。即ち、対象となる AI のタイプや攻撃者の目的によって攻撃を分類するのではなく、その根底にある技術的手法（攻撃メカニズム）に従って分類を行う。対象となる AI のタイプや意図された目的から各攻撃メカニズムを明示的に切り離すことで、本稿を記載の重複を避けるように構成している。そうすることで、記載の意味を把握する上での迷いを減らし、容易な理解と実用的な応用を促すことに焦点を当てている。

表 1 は、本稿における攻撃タイプと NIST の報告書における攻撃タイプとの対応関係を示している。本稿には、I：コードインジェクション攻撃や J：ファインチューニング攻撃など、NIST 報告書ではカバーされていないが注目に値すると我々が考えるいくつかの攻撃も含まれている。本稿では、NIST の徹底的だが重複のあるアプローチを補うために、NIST 報告書の文献レビューと詳細な技術分析を参照しつつも、簡素化した分類と直感的な視覚化を導入している点に特徴がある。

表 1　本稿と NIST AI 100-2e2025 の対応表

本稿での攻撃タイプ	NIST 報告書での攻撃タイプ
§ 5.1 A:モデル抽出攻撃	NISTAML.031 Model Extraction
§ 5.2 B:学習データ情報収集攻撃	NISTAML.033 Membership Inference
§ 5.2.1 B1:メンバーシップ推論	NISTAML.032 Reconstruction
§ 5.2.2 B2:属性推論	NISTAML.034 Property Inference
§ 5.2.3 B3:プロパティ推論	NISTAML.032 Reconstruction
§ 5.2.4 B4:モデルインバージョン	NISTAML.032 Reconstruction
§ 5.2.5 B5:データ再構築	NISTAML.038 Data Extraction

§5.2.6 B6:データ抽出	
§5.3 C:モデルポイズニング攻撃	NISTAML.011,026,051　Model Poisoning
§5.4 D:データポイズニング攻撃	NISTAML.021 Clean-label Backdoor NISTAML.023 Backdoor Poisoning NISTAML.024 Targeted Poisoning
§5.5 E:回避攻撃	NISTAML.022 Evasion NISTAML.025 Black-box Evasion
§5.6 F:スポンジ攻撃	NISTAML.014 Energy-latency
§5.7 G:プロンプト窃盗攻撃	NISTAML.038 Data Extraction
§5.8 H:プロンプトインジェクション攻撃	NISTAML.015 Indirect Prompt Injection NISTAML.018 Prompt Injection NISTAML.027 Misaligned Outputs NISTAML.035 Prompt Extraction NISTAML.036 Leaking Info from User Interactions NISTAML.039 Compromising Connected Resources
§5.9 I:コードインジェクション攻撃	
§5.10 J:ファインチューニング攻撃	
§5.11 K:ロウハンマー攻撃	NISTAML.031 Model Extraction

2.3. MITRE ATLAS™

AI システムに対する敵対的脅威を分類するためのもう一つの関連リソースは、AI システムに対する敵対的攻撃専用の構造化ナレッジベースとして設計された MITRE ATLAS™[57]である。ATLAS は、サイバーセキュリティのコミュニティでよく知られたフレームワークである MITRE ATT&CK®[56]に基づいて、戦術（攻撃者の目標）と技術（それらの目標を達成するために使用される手法）という階層的な概念に従い、現実世界の攻撃シナリオを体系的に分類している。

ATLAS の主な利点はその実践志向であり、攻撃を現実のシステム運用で観測された特定の脅威シナリオに積極的に対応付けている点にある。これは、敵対的な戦術と技術の詳細な説明を文書化された事例にリンクさせることにより、特にサイバーセキュリティ専門家に対して、実用的な洞察を提供する。加えて、ATLAS は攻撃者の能力と目的を効果的に示しており、現実的なリスクアセスメントを促進する。

表 2 は、本稿に示す攻撃タイプと ATLAS との対応関係を示している。本稿は、B2：属性推

論、B3：プロパティ推論、G：プロンプト窃盗攻撃、J：ファインチューニング攻撃、K：ロウハンマー攻撃など、ATLAS に示されていない攻撃についても明らかにしている。つまり、比較的新しい知見を体系的に統合することで ATLAS の実践的な性質を補完し、それによってより広範に新たな脅威を包括している。

表 2　本稿と MITRE ATLAS™の対応関係

本稿での攻撃タイプ	ATLAS での技術（Technique）[57] (Data v4.8.0)
§ 5.1 A:モデル抽出攻撃	AML.T0024.002 Extract ML Model
§ 5.2 B:学習データ情報収集攻撃	AML.T0024 Exfiltration via ML Inference API
§ 5.2.1 B1:メンバーシップ推論	AML.T0024.000 Infer Training Data Membership
§ 5.2.2 B2:属性推論	-
§ 5.2.3 B3:プロパティ推論	-
§ 5.2.4 B4:モデルインバージョン	AML.T0024.001 Invert ML Model
§ 5.2.5 B5:データ再構築	AML.T0024.001 Invert ML Model
§ 5.2.6 B6:データ抽出	AML.T0057 LLM Data Leakage
§ 5.3 C:モデルポイズニング攻撃	AML.T0010.003 ML Supply Chain Compromise: Model AML.T0018.000 Backdoor ML Model: Poison ML Model AML.T0019 Publish Poisoned Datasets AML.T0058 Publish Poisoned Models
§ 5.4 D:データポイズニング攻撃	AML.T0010.002 ML Supply Chain Compromise: Data AML.T0018.000 Backdoor ML Model: Poison ML Model AML.T0020 Poison Training Data AML.T0031 Erode ML Model Integrity AML.T0059 Erode Dataset Integrity
§ 5.5 E:回避攻撃	AML.T0015 Evade ML Model AML.T0031 Erode ML Model Integrity AML.T0043 Craft Adversarial Data
§ 5.6 F:スポンジ攻撃	AML.T0034 Cost Harvesting AML.T0043 Craft Adversarial Data
§ 5.7 G:プロンプト窃盗攻撃	-
§ 5.8 H:プロンプトインジェクション	AML.T0051 LLM Prompt Injection

攻撃	AML.T0054 LLM Jailbreak AML.T0056 Extract LLM System Prompt AML.T0070 RAG Poisoning
§5.9 I:コードインジェクション攻撃	AML.T0011.000 User Execution: Unsafe ML Artifacts AML.T0018.001 Backdoor ML Model: Inject Payload
§5.10 J:ファインチューニング攻撃	-
§5.11 K:ロウハンマー攻撃	-

2.4. OWASP Top 10 for LLM Applications

LLM のセキュリティリスクに取り組んでいるもう一つの影響力のあるリソースは、「OWASP Top 10 for LLM Applications（バージョン 2025）」[94]である。この文書は Open Web Application Security Project（OWASP）によって発行されている。OWASP は Web アプリケーションセキュリティに関する著名な「OWASP Top 10」[88]で最もよく知られ、サイバーセキュリティへの貢献で世界的に認められている国際組織である。

OWASP Top 10 for LLM は、プロンプトインジェクション、機密情報の開示、サプライチェーンの脆弱性、システムプロンプトの漏洩、マルチモーダルインタラクションに関連する新たな懸念など、AI セキュリティの文脈における最も重大な脆弱性を特定している。AI セキュリティの複雑な概念を率直で実用的なガイダンスに効果的に翻訳することでもたらされる実用性と明晰性とに OWASP のアプローチの主な利点がある。

しかし、OWASP のフレームワークは、LLM アプリケーションの当面のセキュリティリスクについて明晰かつ実用的な洞察を提供する一方で、攻撃者の戦術や技術の詳細な分類を提供することを目的とはしていない。その結果、攻撃者に求められる能力や様々な攻撃手法の技術的基盤に関する洞察は限定的である。更に、OWASP は主に LLM 固有の脆弱性に焦点を当てており、より広範な AI システムのアーキテクチャやモダリティへの適用可能性が限定されている可能性がある。

本稿の分類は、より広範な AI モダリティにわたって体系的な構造を提供することで OWASP を補完しており、実践的かつ分析的なユースケースをサポートする。

2.5. 機械学習品質マネジメントガイドライン

更なる関連リソースとして、機械学習品質マネジメントガイドライン（AIQM）[23]がある。AIQM 第 4 版では、特に 10.3 項において、AI セキュリティのカバレッジが大幅に拡大された。AIQM は、機械学習特有の脅威、脆弱性、セキュリティコントロールの体系化に焦点を当てており、学術文献の広範なレビューに基づく本質的な洞察を提供し、論文[43]で技術的な詳細を詳述している。本稿の参考文献の多くが AIQM の参考文献を重なっており、AIQM の貢献に謝意を示す。

AIQM はライフサイクル指向またはプロセス指向の記述スタイルを採用している。このため、AIQM では、システムの影響 、攻撃者の目標、特定の技術など、複数のコンテキストにわたって個々の攻撃の説明が分散しており、深く入れ子になった節を通じて議論を構成している。この構造は、読者が特定の攻撃シナリオについて一貫した理解を得る際の妨げとなるおそれがある。

本稿では対照的に、攻撃のより単純な表現を提供し、詳細だが記載が分散してしまう AIQM のアプローチを補完する。

2.6. その他のガイドラインとフレームワーク

他にも様々な組織が AI システムのリスク管理や安全な開発プラクティスに取り組むガイドラインやフレームワークを公開しているが[1][14][15][62][64][65]、AI に対する特定の攻撃方法の議論は一般にこれらの文書における議論の対象にはなっていない。

最後に、もう一つ触れておくべき関連資料は、J-AISI によって発行された「AI セーフティに関するレッドチーミング手法ガイド」[38]である。このガイドは J-AISI における草創期の取り組みのひとつとして、LLM システムを対象としたレッドチーム演習の計画・準備・実施のための実践的かつ構造化された推奨事項を提供する。本ガイドでは、実践的なシステム評価とリスク軽減を支援する具体例として、プロンプトインジェクションなどの代表的な攻撃が簡潔に紹介されている。しかし、このガイドの主な目的は、あくまでもレッドチーミングのプロセス全体を概説することにある。したがって、敵対的脅威のカバレッジは限定的であり、詳細な説明も攻撃の包括的な分類も提供していない。本稿は、レッドチーミング手法ガイドよりも深い技術的説明、より広範な敵対的脅威、そして攻撃メカニズムとセキュリティ上の影響に関する詳細な洞察を提供することで、このガイドを補完する。本稿の提示する分類は、敵対的脅威を特定し優先順位付けするための手頃な情報源として機能し、J-AISI が提案する実践的なレッドチーミング方法論を強化する。

3. AI システムおよびスコープ

後続の議論の土台を築くため、本節では AI システムの一般的なアーキテクチャを定義し、本稿で検討する攻撃のスコープを概説する。

3.1. AI System Architecture AI システムのアーキテクチャ

AI システムとは、AI モデルをコンポーネントとして組み込んだ情報システムである。図 2 は、本稿で想定する一般的な AI システムのアーキテクチャを示している。このシステムは、概念的に 2 つの主要な環境に分かれている：

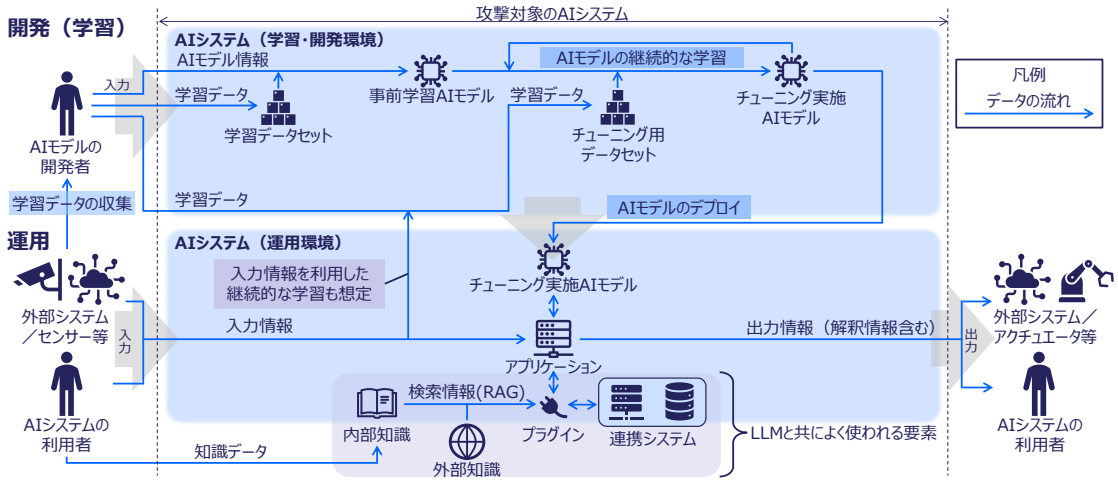

図 2　想定する AI システムアーキテクチャ。この図は、モデルの学習からデプロイ、そして継続的な運用に至るシームレスなデータの流れを示している。学習環境では、AI モデルの開発者から提供される学習データだけでなく、運用過程で得られたデータを活用した、継続的または定期的なファインチューニングが行われることを想定している。運用環境では、情報検索のための内部知識や外部知識、およびプラグインを介して接続された連携システムの使用を想定している。これらの要素は LLM と共に使われることが多いが、AI モデルは LLM に限定されない。

- **学習環境（図 2 の上部）**：学習環境において、実世界の運用中に収集されたデータを含むデータセットを使用し、AI モデルに対する学習およびファインチューニングが行われる。運用過程で得られたデータを活用した、継続的または定期的なファインチューニングのシナリオも本稿の考慮の対象とする。
- **運用環境（図 2 の下部）**：学習済みの AI モデルは運用環境にデプロイされ、アプリケーションフレームワークに統合される。外部システム、センサー、そして人間のユーザーからの入力データがモデルによって処理される。システムが生成する出力の宛先は、デ

ータを入力したのと同一のエンティティだけではなく、異なる外部エンティティになる可能性がある。

本稿の図において、データは基本的に左から右へ流れるものとする。図 2 に示される各コンポーネントの詳細な説明は、付録 A.1 で提供される。図示したシステムアーキテクチャでは、主に LLM に関連するプラグインや検索拡張生成（RAG）を介した内部知識と外部知識の統合を考慮している。しかし、本稿の議論は LLM ベースのシステムに限らず適用されることに注意されたい。

この抽象モデルは、複雑な現実世界のアーキテクチャを簡素化しつつも、脆弱性を現実的かつ実践的に分析するのに十分な詳細を保持している。同時に、これは万能のフレームワークではない。例えば連合学習のような特定の高度な機械学習アーキテクチャはこの抽象化に適合せず、固有の脆弱性についても個別の分析が必要となる。

3.2. 本稿で扱う攻撃のスコープ

AI システムのセキュリティ確保にあたっては従来のサイバー脅威を軽減するための標準的なサイバーセキュリティ対策が本質的に含まれる。本稿では、不正なネットワークアクセスからの保護といった一般的なサイバーセキュリティ対策はすでに実施されている前提とする。その上で本稿では、*AI モデルの学習または推論への介入、ならびに標的となるシステム上の AI モデルの入出力の改ざんまたは悪用を伴う、AI に特有の攻撃*に焦点を当てる。

標準的なサイバーセキュリティの脆弱性を悪用する攻撃については、例えば AI モデルの窃取につながる不正アクセスであっても、その攻撃が AI システムに固有のプロセスを狙って介入するものでない限り、本稿では取り上げない。

AI システムに対する攻撃の研究は大幅に増えており、既存のすべての文献を網羅的に調査することは現実的でなく、本稿もその意味での完全性を主張することはない。その代わりに本稿では、第 2 節で概説したようなよく知られた基礎的研究と、学術会議での発表に見られる最近の進展の両方を参照する。

4. 攻撃タイプとセキュリティ上の影響

本節では、AI システムを標的とする攻撃の本稿における分類を説明し、これらの攻撃および攻撃がもたらす AI システムへのセキュリティ上の影響との関係を解説する。

4.1. 本稿のアプローチ

本稿で示される攻撃と影響のタイプは、第 2 節で議論した情報源のレビューに基づき、以下の手順に従って導出された。

(1) **特定・抽出**：第 2 節で参照した資料から、潜在的な攻撃と影響に対応する関連記述を特定・抽出した。また、ACM CCS などの国際会議における最近の論文[18][46][60][100][103]からも、第 2 節の参照資料にない攻撃手法を抽出した。
(2) **AI システムアーキテクチャへのマッピング**：上記で抽出した攻撃手法を、攻撃対象領域や対象コンポーネントといった技術的観点を考慮しながら、図 2 の AI システムアーキテクチャに対応付けた。
(3) **グループ化**：(2)のマッピングの結果を用い、類似または関連する攻撃を攻撃タイプにグループ化した（図 1 参照）。(1)で抽出された影響も、AI システムアーキテクチャ内で影響を受けるコンポーネントと CIA の 3 要件の観点に基づきグループ化した（図 3 参照）。
(4) **攻撃と影響の関連付け**：参照資料に基づき、各攻撃タイプを AI システムへのセキュリティ上の影響に関連付けた（図 4 および表 4 参照）。

以下の項では、このアプローチによって得られた攻撃と影響の最終的な分類について説明する。

4.2. AI システムに対する攻撃の分類

図 1 が示すように、本稿では AI システムに対する攻撃を A から K までの 11 のタイプに分類する。各タイプは、AI システムにおける AI 特有の脆弱性に関連付けられた固有の攻撃ベクトルを表している。各攻撃タイプの詳細な説明は第 5 節で与える。

各攻撃に対し本稿で採用した命名規則は、学術界および産業界の確立された用語体系に概ね準拠しており、それによって可読性と理解を高めている。ただし、本稿が示す攻撃カテゴリの中で、学習データ情報収集攻撃（攻撃 B）については、モデルの学習データセットに関する情報を抽出または推論することを意図した攻撃を包含するクラスとして新たに導入したものである。

メカニズム上は同一の攻撃を異なる観点から説明する複数の確立された用語がある場合には、本稿では主としてその攻撃の技術的メカニズムに基づいた用語を選択している。例えば、「プロンプトインジェクション」と「ジェイルブレイク」は技術的メカニズムとしては

同じ攻撃を指すことが多いが、本稿ではこれを指す言葉として「プロンプトインジェクション」を採用している。なぜなら、この言葉が根底にあるメカニズムを反映しているのに対し、「ジェイルブレイク」はセーフガード回避という攻撃の目的や結果を強調するものだからである。命名に関する本稿の規則は、単一の攻撃技術が複数の敵対的目的を果たし得ることを認識することにより、正確な理解を提供することを狙っている。本稿において攻撃タイプとセキュリティ上の影響を分離した分類体系を示しているのは、ひとつの攻撃技法が多種多様な目的・意図で利用されるという観察に基づいている。

なお、幾つかの攻撃については、普及済みの用語が攻撃者の方法の技術的側面を適切に表現しているとは言えず、確立された代替表現も存在しない。このような場合には、読者にとっての馴染みやすさを確保しつつ既存文献との互換性を保つために、新たな用語を導入せずに既存の用語をそのまま利用している。

4.3. 攻撃タイプと AI システムへの影響の関係

図 1 は AI システムに対する様々な攻撃の概要を示し、図 3 はこれらの攻撃が AI システムに及ぼし得るセキュリティ上の影響を示している。全体として、モデル漏洩や学習データ漏洩から、運用上の不具合、セーフガード回避、システム侵害に至るまで、本稿では 11 タイプの影響が特定されている。付録 A.2 では、これら 11 の影響について追加の説明を提供している。

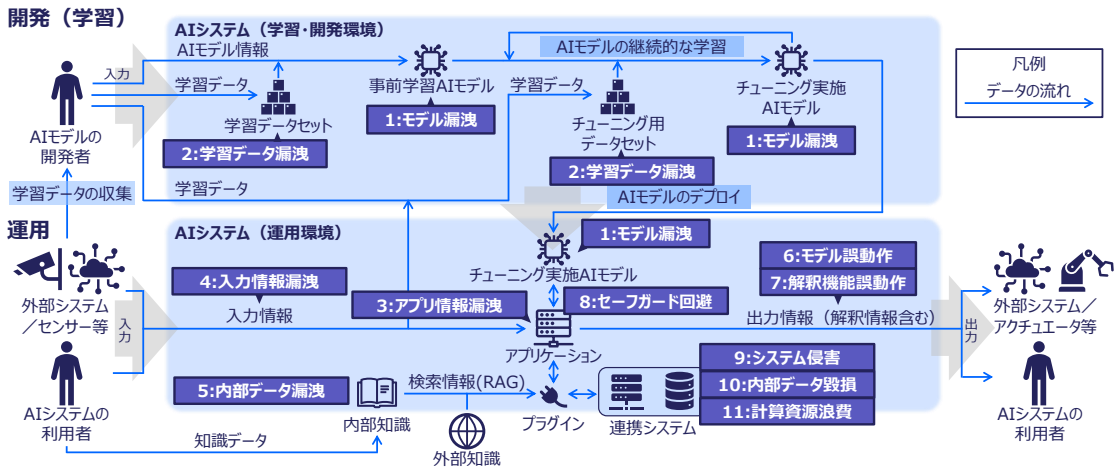

図 3　図 1 に示した AI システムに対する攻撃による影響（モデル漏洩のように同じ種類の影響が複数の箇所に生じる場合もある）

図 4 は、情報セキュリティにおける CIA の 3 要件として知られる機密性・完全性・可用性の観点から、攻撃タイプ（図 1）とそれらが AI システムへ及ぼす影響（図 3）との関係を視

覚的に対応付けている。破線の矢印は後続の攻撃を容易にするおそれのある影響を示しており、潜在的な連鎖的リスクを捕捉する一助とすることを意図している。特筆すべきことのひとつは、セキュリティ上の影響として「9：システム侵害」を強調している点である。システム侵害は、従来のサイバーセキュリティの脅威を含む様々な後続の攻撃を可能にし、CIA の 3 要件のすべてに影響する可能性があるためである。

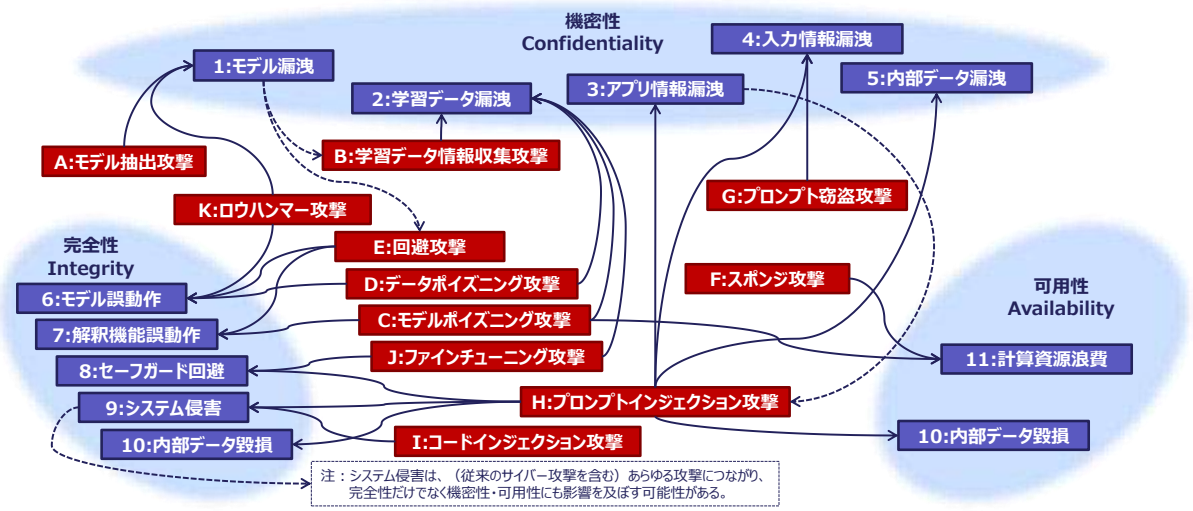

図 4 攻撃とその影響の関係（破線は矢印の先の攻撃に利用される可能性を示す）

表 3 は、AI システムへの影響から攻撃タイプを特定するための逆引きの手段を提供する。表にまとめた各攻撃タイプは、第 5 節の詳細な議論にハイパーリンクされている。この表は、想定される影響と対処が必要な関連攻撃タイプとを直接結び付けており、リスクアセスメントやレッドチーミングに活用できる。

表 3　攻撃とその影響に関する逆引き表

C	I	A	影響	攻撃タイプ	節・項
○			1:モデル漏洩	A:モデル抽出攻撃	§5.1
○				K:ロウハンマー攻撃	§5.11
○			2:学習データ漏洩	B:学習データ情報収集攻撃	§5.2
○			3:アプリ情報漏洩	H:プロンプトインジェクション攻撃	§5.8
○			4:入力情報漏洩	G:プロンプト窃盗攻撃	§5.7
○				H:プロンプトインジェクション攻撃	§5.8
○			5:内部データ漏洩	H:プロンプトインジェクション攻撃	§5.8
	○		6:モデル誤動作	D:データポイズニング攻撃	§5.4

	○			E:回避攻撃	§5.5
	○			K:ロウハンマー攻撃	§5.11
	○		7:解釈機能誤動作	C:モデルポイズニング攻撃	§5.3
	○			E:回避攻撃	§5.5
	○		8:セーフガード回避	H:プロンプトインジェクション攻撃	§5.8
	○			J:ファインチューニング攻撃	§5.10
	○		9:システム侵害	H:プロンプトインジェクション攻撃	§5.8
	○			I:コードインジェクション攻撃	§5.9
	○	○	10:内部データ毀損	H:プロンプトインジェクション攻撃	§5.8
		○	11:計算資源浪費	C:モデルポイズニング攻撃	§5.3
		○		F:スポンジ攻撃	§5.6

5. 各攻撃タイプのメカニズム

本節では、図 1 で紹介した 11 の攻撃タイプ（A〜K）について詳細に説明する。各攻撃タイプについて、攻撃の開始からシステムへの最終的な影響に至るまでの全体の流れを図示する。

表 4 は、既知の攻撃タイプとそれらが AI システムに及ぼすセキュリティ上の影響をまとめたものである。各攻撃タイプについて、想定される攻撃の前提条件が、関連する論文への参照とともに、文献で実証されているモダリティと AI モデルのタイプと共にリストされている。

表 4 に記載されているモダリティと対象となる AI モデルのタイプは、文献に記載されている実験的検証に基づく要約であって、必ずしも攻撃手法自体の本質的な限界を示すものではないことに留意する必要がある。

表 4　AI システムに対する既知の攻撃と影響
（これは文献に基づく要約であって、各攻撃に関わるモダリティや対象を記載の範囲に限定する意図はない。ID と略語は第 5 節の冒頭で定義されている。）

攻撃タイプ		
影響	攻撃の前提条件	モダリティおよび AI モデルタイプ
§5.1 A:モデル抽出攻撃		
1:モデル漏洩	モデルへの大量クエ	Image: NN [20][42][67][68], LR

	リー	[87] Tabular: RR [92], DT [87], RL, NN, SVM [87][92]
§5.2.1 B1:メンバーシップ推論		
2:学習データ漏洩	モデルへの入力	Image, Tabular: LR, DT [99], NN [2][99] Text: NN [2]
	モデルへの大量クエリー	Image, Tabular: NN [51][80] Text: LM [52]
	モデルの内部情報へのアクセス	Image, Tabular: NN [59]
§5.2.2 B2:属性推論		
2:学習データ漏洩	モデルへの入力	Image, Tabular: LR, DT [99]
	モデルの内部情報へのアクセス	Image, Text: NN [84]
§5.2.3 B3:プロパティ推論		
2:学習データ漏洩	モデルの内部情報へのアクセス	Image, Tabular: NN [29] Audio: NN, SVM, HMM, DT [4] Network Traffic: NN, SVM, HMM, DT [4]
§5.2.4 B4:モデルインバージョン		
2:学習データ漏洩	モデルへの大量クエリー	Image, Tabular: DT, NN [28]
	モデルの内部情報へのアクセス	Text: Transformer [101]
§5.2.5 B5:データ再構築		
2:学習データ漏洩	モデルの内部情報へのアクセス	Image: NN [5][7][34]
	モデルへの大量クエリー	Text: LM [52]
§5.2.6 B6:データ抽出		
2:学習データ漏洩	モデルへの大量クエリー	Text: LSTM, RNN [11], LM [13][52] Text to Image: Diffusion [10]
§5.3 C:モデルポイズニング攻撃		

2:学習データ漏洩	学習プログラムの改竄	Image: NN [83] Text: SVM, LR [83]
7:解釈機能誤動作	モデルの改竄	Tabular: NN, EM [82]
11:計算資源浪費	モデルの改竄	Image: NN [19]
§5.4 D:データポイズニング攻撃		
2:学習データ漏洩	学習データへのインジェクション＋モデルへの大量クエリー	Tabular: LR, NN [54], Text: LR [54]
6:モデル誤動作	学習データへのインジェクション	Image: NN [9] Text: LM [76] Tabular: Linear Regression [41]
§5.5 E:回避攻撃		
6:モデル誤動作	モデルへの大量クエリー	Image: NN, LR, SVM, DT, kNN [71] Text: LM [6]
	モデルの内部情報へのアクセス	Image: NN [30][86] Text: LSTM [26]
	モデルへの入力	Text: Transformer [33], LSTM, DA [91]
7:解釈機能誤動作	モデルへの大量クエリー	Image: NN [24]
§5.6 F:スポンジ攻撃		
11:計算資源浪費	モデルへの入力	Image, Text: NN [81]
	モデルへの大量クエリー	Text: LM [6]
§5.7 G:プロンプト窃盗攻撃		
4:入力データ漏洩	モデルの出力へのアクセス	Text to Image: Diffusion [79]
§5.8 H:プロンプトインジェクション攻撃		
4:入力データ漏洩	ユーザー参照情報に対するポイズニング User-referenced info poisoning	Text: A Specific Chat Service [75]
3:アプリ情報漏洩	（システム API 経由での）モデルへの入	Text: LM [102]

	力	
5:内部データ漏洩 10:内部データ毀損	（システム API 経由での）モデルへの入力	Text to Tabular: LM [72]
8:セーフガード回避	モデルへの入力	Text: LM [16][49][55][73][78][91][93][106]
	モデルの内部情報へのアクセス	Text, Image: NN [12]
9:システム侵害	AI システム参照情報に対するポイズニング AI system-referenced info poisoning	Text, Image to Text: LM [32]
§5.9 I:コードインジェクション攻撃		
9:システム侵害	悪意あるモデルの配布	Any: Any [103]
§5.10 J:ファインチューニング攻撃		
2:学習データ漏洩	ファインチューニング機能へのアクセス	Text: LM [18]
8:セーフガード回避	モデルの内部情報へのアクセス	Text: LM [100]
§5.11 K:ロウハンマー攻撃		
1:モデル漏洩	物理メモリアクセス	Image: NN [74]
6:モデル誤動作	物理メモリアクセス	Image: NN [46]
	モデルの内部情報へのアクセス＋物理メモリアクセス	Image, Text: LM, Transformer [60]

注：表 4 では、AI モデルのタイプとして以下の略語を使用している：NN（ニューラルネットワーク）、LR（ロジスティック回帰）、RR（リッジ回帰）、SVM（サポートベクターマシン）、DT（決定木）、KNN（k 近傍法）、HMM（隠れマルコフモデル）、LM（言語モデル）、LSTM（Long-Short Term Memory）、RNN（リカレントニューラルネットワーク）、EM（アンサンブル法）、DA（Decomposable Attention）。

5.1. A：モデル抽出攻撃

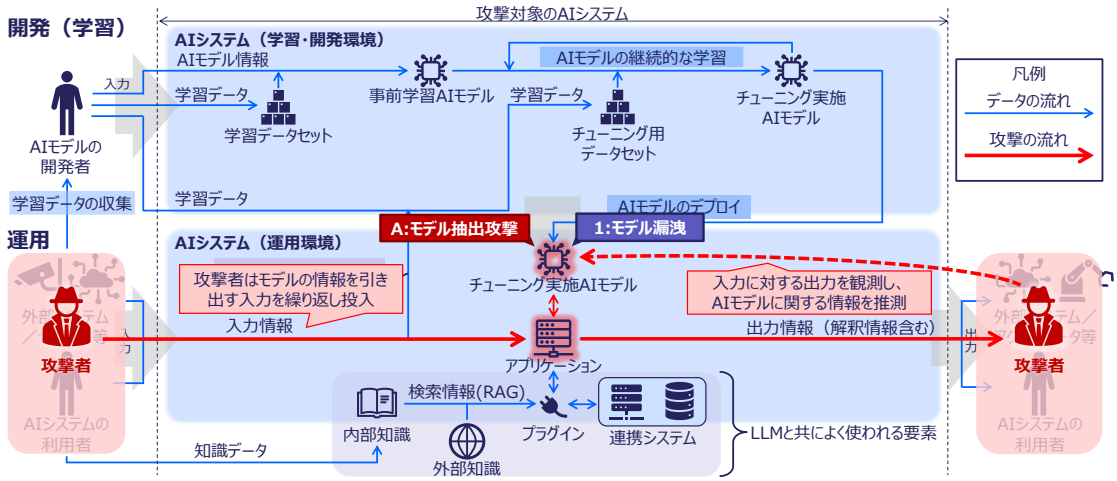

図 5　A：モデル抽出攻撃（概要）

モデル抽出攻撃は、AI モデルの内部詳細に直接アクセスできない攻撃者（例：公開された API を介してやり取りする利用者）が実行する。様々な入力をモデルに体系的に問い合わせ、対応する出力を分析することによって、攻撃者は基盤となるモデル構造やパラメータに関する機密情報を推測しようと試みる（図 5）。

モデル抽出の動機は、AI の経済的・社会的重要性が増すにつれて高まっている。通常、最先端の AI モデルは計算資源と独自のデータセットの両面で、学習への大規模な投資を行った結果である。例えば抽出した AI モデルを闇市場で販売するような攻撃者にとって、モデル抽出の試みは非常に価値のあるものとなり得る。モデル抽出は、5.2 項の学習データ情報収集攻撃（攻撃 B）や 5.5 項の回避攻撃（攻撃 E）といった、より高度な攻撃の足掛かりとしても機能することが多い。

以下のように様々な形態のモデル固有の情報を抽出できることが先行研究において実証されている：

- モデルアーキテクチャ（例：ニューラルネットワークの構造と構成）[67]
- ハイパーパラメータ（例：学習時の過学習を防ぐための正則化パラメータ）[92]
- パラメータ（例：ニューラルネットワーク内のニューロンの重み）[87]
- 決定境界（入力空間でモデルの予測クラスが変化する境界）[42]
- 機能性（内部構造なしのモデルの入出力関係）[20][68]

表 4 は、本稿で取り上げた代表的な研究によって調査されたモダリティと AI モデルのタイプをまとめている。

モデル抽出攻撃を軽減するいくつかのアプローチが提案されている。防御の基本原則は、関連する信頼度スコアなしで離散的な分類結果のみを提供するなどの方法で、不要な情報開示を最小限に抑えることである。PRADA[42]などのツールを含む検知指向の手法は、クエリーパターンを監視して不審な抽出の試みを特定する。また、正規の利用を大幅に損なわない範囲で、入力に摂動を加える手法や難読化する技術が、モデル情報の正確な推測を試みる攻撃者を妨害するために提案されている[31]。AI モデルの抽出攻撃に対する脆弱性を評価する ML-Doctor[50]のようなフレームワークも提案されている。アンサンブル学習もまた、モデルの挙動を難読化または多様化することによってロバスト性を提供することができ、抽出プロセスを複雑にし得る。

攻撃自体を防ぐものではないが、事後的な対策として透かし[48]とフィンガープリンティング[8]の手法が研究されている。これらは、抽出攻撃の後に、モデルの不正なコピーを検出または証明するメカニズムを提供する。

5.1.1. モデル抽出と知識蒸留

モデル抽出と、よく知られた機械学習の手法である知識蒸留との間には、概念的な類似性がある一方で決定的な違いがある。**知識蒸留**とは、大規模で複雑な（「教師」）モデルの挙動を模倣する相対的に小規模で計算効率の高い（「生徒」）モデルを作成するために、開発者が意図的に使用する技術を指す。両者の根本的な違いは意図と正当性にある。知識蒸留は効率とスケーラビリティの目的のために、正規の主体によって明示的に意図・実施されるものであって、知的財産の不正な複製や再配布を狙ったものではない。このため、知識蒸留を敵対的攻撃に分類することは通常ない。他方で、不正なモデルの複製に関する懸念から、多くの AI サービス提供者は利用規約で知識蒸留を明示的に禁止している。

5.2. B：学習データ情報収集攻撃

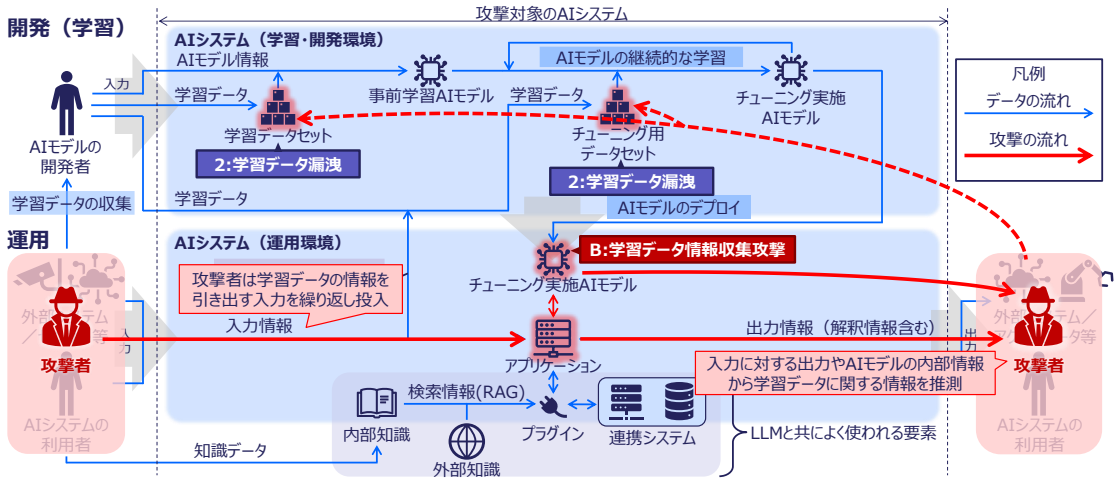

図 6　B：学習データ情報収集攻撃（概要）

直接的なデータ窃取とは対照的に、**学習データ情報収集**と呼ばれるカテゴリの攻撃は、AI モデルの入力とそれに対応する出力を体系的に監視し、場合によってはモデル自体に関する内部情報を悪用することによって、AI モデルの学習データに関連する情報を抽出または推論することを伴う（図 6 参照）。

ここでの学習データ情報という用語は、元の学習サンプルの直接的な抽出にとどまらず、学習データに関する間接的な情報が明らかになる広範なシナリオを包含する。典型的な攻撃には次のものが含まれる：①特定の個人や記録が学習データセットの一部であったかどうかを推論する**メンバーシップ推論**。②データセット内の特定のレコードに関する未知かつ機微な属性を推論する**属性推論**。③データセットの一般的な統計的特性を推論することを目指す**プロパティ推論**。④モデルの出力を繰り返し問い合わせることで代表的なクラスレベルのサンプルを再構築する**モデルインバージョン**。⑤攻撃者がモデルの内部パラメータや中間表現を利用して理論的には特定の学習データサンプルを精緻に再構築する**データ再構築**。⑥そして特にデータ生成を想定して、AI に記憶された正確な学習サンプルをモデルの出力から直接取得する**データ抽出**。

以下の項で、各攻撃タイプについて詳細に説明する。

5.2.1. B1：メンバーシップ推論

この攻撃では、モデルの入出力の挙動を監視することによって、特定のデータサンプルが学習データセットに含まれていたかどうかを推論する。特にモデルが機密データで学習されて

いる場合、このような攻撃は深刻なプライバシー上の脅威をもたらす。例えば、ローン不履行の財務記録で学習したAIシステムでは、特定の個人の履歴がその学習データセットの一部であったかどうかを漏洩させるだけで、機密性の高い個人情報を明らかにしてしまう可能性がある。

この攻撃の典型的な実装では、ターゲットモデルの挙動を模倣するシャドウモデルを使用する。攻撃者はまず、各サンプルのメンバーシップ状態（含まれているか、含まれていないか）が既知であるデータセットを使用して、シャドウモデルを学習させる。次に、別のモデル（しばしば攻撃モデルと呼ばれる）が、シャドウモデルの学習データセットに含まれるデータサンプルと含まれないデータサンプルに対するシャドウモデルの挙動の違いを学習する。この攻撃モデルを使用して、攻撃者はターゲットモデルの出力から特定のデータポイントのメンバーシップ状態を推測できる。表4は、様々なモダリティとモデルタイプにわたるメンバーシップ推論攻撃に関する文献をまとめている。

メンバーシップ推論攻撃のリスクを緩和するいくつかの対策が提案されている：

- 学習環境において
  - 過学習の防止：モデル学習時に正則化を適用し、学習データの意図しない記憶を最小限に抑える。
  - 評価ツール：専門の分析ツール（例：ML Privacy Meter[58]またはML-Doctor[50]）を活用し、メンバーシップ推論攻撃を積極的に評価・軽減する。
- 運用環境において
  - 出力制御：出力の詳細度を低減する。信頼度スコアなしで上位の予測のみを報告したり、出力の数値精度を下げたりするなどの例がある。
  - 差分プライバシー：差分プライバシーのメカニズムを組み込み、個々のデータポイントに関する情報漏洩を数学的に制限する。

5.2.2. B2：属性推論

この攻撃では、公開または部分的に既知の情報（例：ソーシャルメディアから得られる年齢、性別）とモデルの入出力の挙動を悪用することにより、AIモデルの学習に使用されたデータサンプルの未知かつ機微な属性（例：特定の遺伝子型）を推論する [99]。即ち、攻撃者は特定の属性に関する部分的な知識を用いて、機密であるべき属性を推論することができ、非常にプライベートな個人情報が意図せず開示される可能性がある。表4は、様々なモダリティとモデルタイプにわたる属性推論攻撃に関する文献をまとめている。

属性推論攻撃のリスクを緩和するにはメンバーシップ推論攻撃の場合と同様の対策が有効である。具体的には、過学習を防ぐための正則化や、機密属性の漏洩を制限するための差分プライバシー技術の導入などがある。

5.2.3. B3：プロパティ推論

この攻撃では、モデル出力から得た情報を分析することにより、学習データセットの全体的な統計的特性を推論する。特定のデータサンプルや属性を対象とするメンバーシップ推論や属性推論とは対照的に、プロパティ推論攻撃は、データセットレベルの一般的な特性（例：分布特性、特定の人口統計グループの割合）を対象とする。

基盤を築いた論文[29]の図 1 は、プロパティ推論攻撃がどのように実行され得るかの概要を示している。まず、攻撃者は複数のシャドウ分類器を学習させるが、その際、ある統計的特性 P を持つデータセットで学習させるものと、その特性を持たない（$\bar{P}$と表記される）データセットで学習させるものを用意する。次に、内部パラメータや出力といった特徴量がこれらのシャドウ分類器から抽出され、メタ学習データセットを構築するために使用される。続いてこのデータセットを、特性 P 有りで学習されたモデルと P 無しで学習されたモデルとを識別可能なメタ分類器を学習させるために使用する。攻撃者は最終的に、シャドウ分類器から抽出したのと同様の特徴量をターゲットモデルから抽出することによって、ターゲットモデルの学習データセットが特性 P を所有しているかどうかを推論できる。表 4 は、様々なモダリティとモデルタイプにわたるプロパティ推論攻撃に関する文献をまとめている。

プロパティ推論攻撃のリスクを緩和する戦略のひとつとしては、学習データセットにノイズデータを導入する方法が提案されている[29]。

5.2.4. B4：モデルインバージョン

この攻撃では、モデルの特定のクラス予測を強力に活性化する入力を生成することで、特定のクラスの代表的なデータを再構築する。攻撃者は通常、モデルのフィードバック（例：信頼度スコア）を参照しながら、ターゲットクラスに対する分類器の信頼度を最大化するように入力データを繰り返し調整し、そのクラスを特徴づけるデータへと収束させる。端緒となった論文[28]は、論文中の図 1 で攻撃例を示しており、攻撃によって再構築された画像と、モデルの学習データセットからの対応する元の画像の両方を図示している。再構築の成功は条件に大きく左右されるものの、この攻撃は潜在的なプライバシーリスクを浮き彫りにしている。

表 4 は、モデルインバージョン攻撃に関連する文献をまとめている。

該当文献で言及されているシンプルな対策は、モデルによって返される信頼度スコアをマスキングまたは制限することである。

5.2.5. B5：データ再構築

この攻撃は、学習済み AI モデルの内部表現や出力から、理論的には特定のデータサンプルを再構築できる。あるクラスの一般的な表現を生成するモデルインバージョンとは対照的に、データ再構築攻撃は、元の学習サンプルに酷似した、具体的かつ特定のサンプルを再構築する。

モデルのパラメータや出力を利用して個々の学習サンプルを正確に再構築する方法を実証した文献の詳細は表 4 にまとめられている。

この種の攻撃は、学習データセットに含まれる機密情報や個人情報が正確な形で暴露され、機密性や規制遵守に違反する可能性を具体的に示しており、プライバシー上の重大な懸念を引き起こす。典型的な対策には、差分プライバシー技術の導入、出力のランダム化、および、モデルの中間表現へのアクセスの厳格な制御が含まれる。

5.2.6. B6：データ抽出

この攻撃は、特に言語モデルや拡散モデルといった生成 AI システムにおいて、学習済みモデルの出力から、記憶された正確な学習データを特定・取得できるよう特別に設計されている。データ抽出では、モデルへの反復的なクエリーを通じて、生成 AI 内に記憶されたデータサンプルを抽出する。その際には、意図されない記憶現象を悪用するのが典型的なアプローチである。これは、一般的なクラス表現を対象とするモデルインバージョンとも、また、（モデルの出力ではなく）内部表現から特定のデータサンプルを理論的に導出するデータ再構築とも対照的である。

注目すべき例は、Carlini ら[13]によって実証された LLM に対する攻撃である。彼らの手法は、個人情報や機密情報の直前に現れる確率が高い初期プロンプトを与え、多数の（後続する）候補テキストの補完を生成した後、メンバーシップ推論技術を使用して、これらの補完のうちどれが実際に学習データに出現したかを特定することを伴う。

表 4 は、データ抽出攻撃に関する文献をまとめている。

文献で議論されているリスク緩和戦略には、意図しない記憶を減らすための差分プライバシー技術の導入や正則化技術の使用が含まれる。

5.3. C：モデルポイズニング攻撃

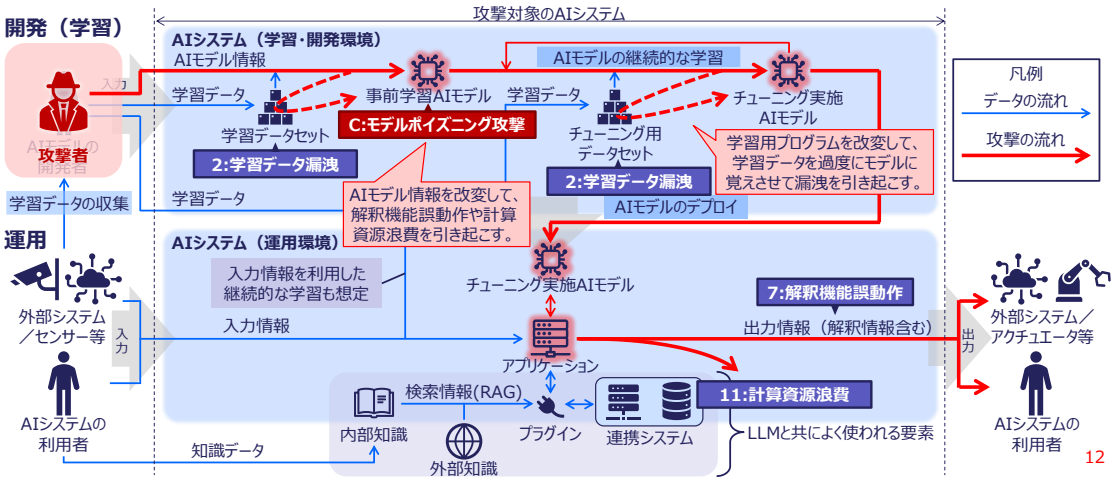

図 7　C：モデルポイズニング攻撃（概要）

モデルポイズニング攻撃は、AI モデル自体またはその学習用プログラムを直接操作し、モデルの挙動を侵害する。学習データは変更しない。悪意のある入力がデータセットに導入されるデータポイズニングとは対照的に、モデルポイズニングは、モデルの基盤となる構造、パラメータ、または学習プロセスを改変する。

先行研究[83]で論じられたモデルポイズニングの典型例は、悪意ある攻撃者が学習アルゴリズムや設定パラメータを改変し、学習データの過度の記憶を誘発するものである。過剰に強化された記憶は、推論時に学習時の姿そのままに再現される可能性が高いため、機密性の高い学習データの漏洩リスクを大幅に高める。このような攻撃は最終的に、AI 動作の解釈機能の深刻な誤動作、非効率なモデル性能による過度の計算資源浪費、または運用時に外部送信するクエリーとしてのデータ漏洩につながる可能性がある（図 7 参照）。

具体的な実験的実証、攻撃の能力、および様々なモダリティとモデルタイプへの影響の概要は表 4 で詳述している。

モデルポイズニング攻撃によって引き起こされるリスクを軽減するには、モデルと学習手順の両方に対する堅牢な検証プロセスが不可欠である。検証プロセスを構成する戦略には、厳格なコード監査、モデルの完全性チェック、安全で耐タンパー性のあるハードウェアまたはソフトウェアフレームワークの採用が含まれる。更に、暗号技術を用いた完全性検証手法や、AI 成果物の安全で追跡可能なバージョン管理を併用することで、攻撃対象領域を効果的に削減できる。

5.4. D：データポイズニング攻撃

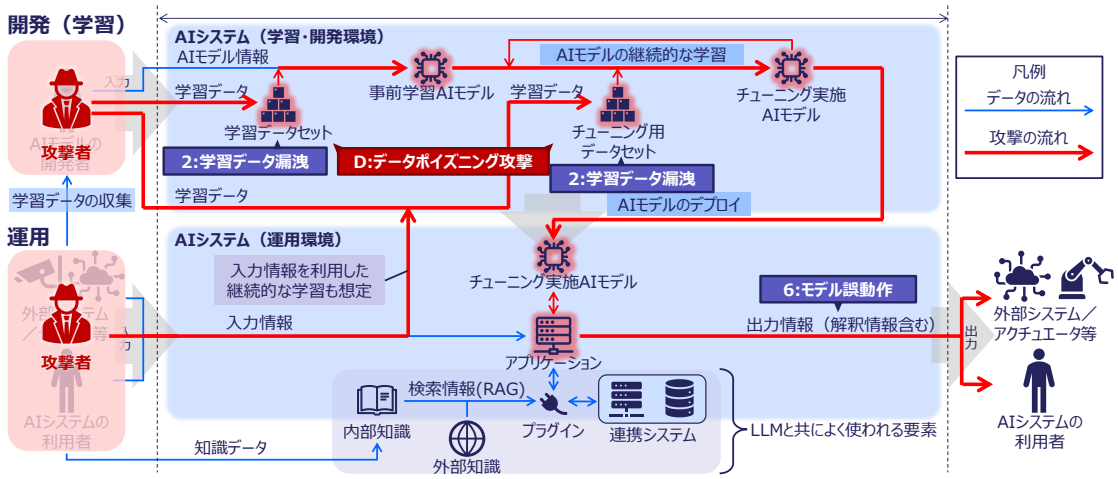

図 8　D：データポイズニング攻撃（概要）

データポイズニング攻撃では、モデルの推論精度を損なう、あるいは推論時に特定の誤動作を誘発することを目的として、AI モデルの学習データセットに対し悪意のあるデータを意図的に混入する。図 8 は、データポイズニング攻撃の全体の運用フローを示しており、悪意のあるデータが学習環境に挿入されることで、運用時のモデルの挙動にどのように影響を与えるかを示している。大まかに言えば、これらの攻撃は主に 3 つのタイプに分類できる。

- **標的型ポイズニング**: これらの攻撃は、特定の入力やデータポイントに対するモデルの挙動を変化させ、特定の標的サンプルに対するモデルの予測を誤らせる。
- **非標的型ポイズニング**: 標的型攻撃とは対照的に、非標的型ポイズニングはモデルの全体的な精度を無差別に低下させる。攻撃者は、学習データセット全体に広範に誤解を招くデータやノイズデータを導入することで、モデルのパフォーマンス全般を低下させる。すなわち、様々な入力に対し予測の信頼性を低下させる。
- **バックドアポイズニング**: バックドアポイズニングでは、選ばれた学習サンプルに巧妙なトリガーを埋め込むことで、推論時にこれらのトリガーが存在する場合にのみ、モデルに不正な出力を生成させる。言い換えれば、ポイズニングされたモデルは、無害な入力に対しては正常に動作するが、トリガーとなる入力に遭遇すると攻撃者の意図に沿って正常動作から逸脱する。

バックドアポイズニングは、特定の入力に対するその条件的な性質から、標的型ポイズニングの特殊なサブセットと見なすことができる。しかし、トリガーメカニズムを必要とするという特徴的な要件のために、しばしば別個に扱われる。

5.4.1. バックドアポイズニングの詳細な分類

既存のサーベイ[47]は、バックドアポイズニング攻撃を更に以下のように分類している（図9参照）。

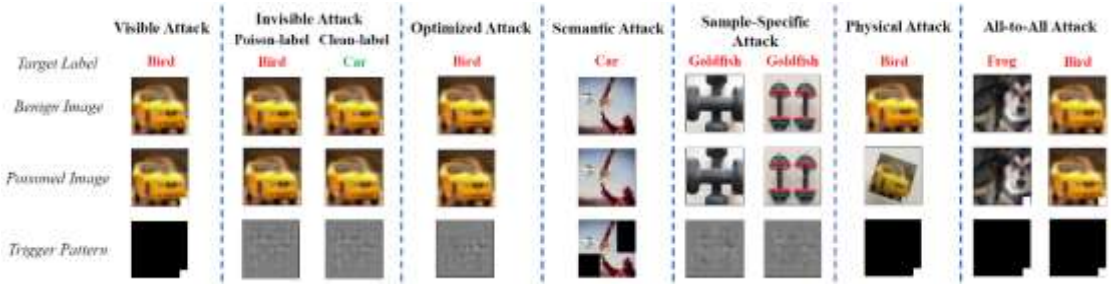

図9　異なるタイプのバックドアポイズニングによって生成されたポイズニング結果の例。Yiming Li, Yong Jiang, Zhifeng Li, Shu-Tao Xia による"Backdoor Learning: A Survey" [47]（arXiv:2007.08745 [cs.CR]）の図4からの引用［CC BY 4.0 ライセンス］

(1) **可視的攻撃**：明示的に視認可能なトリガー（例：画像の隅にある白い四角のような明確なマーカー）を使用し、モデルにトリガーと不正なラベルを関連付けさせる。
(2) **不可視的攻撃**：巧妙で人の目では分からない摂動をトリガーとして組み込み、検出をより困難にする。
(3) **最適化攻撃**：ポイズニングの効果を最大化するために、高度な最適化手法（例：普遍的敵対的摂動）を利用する。
(4) **セマンティック攻撃**：視覚的な摂動ではなく特定の意味論的な組み合わせに依存し、意味論的にトリガーされた入力（例：「鳥」と「人間」の両方を含む画像が「車」として誤分類される）をモデルに誤分類させる。
(5) **サンプル固有攻撃**：特定のポイズニングされたサンプルに合わせて個別に調整された、固有のトリガーパターンを伴う。
(6) **物理攻撃**：センサーによって把握される物理世界においてポイズニングパターンを実装し、現実世界での適用可能性を実証する。
(7) **All-to-All 攻撃**：従来の all-to-one バックドア攻撃（訳注：ポイズニングされた複数のサンプル（All）に単一のラベル（One）を割り当てるタイプのバックドア攻撃）と比較して、標的指向の防御策の回避を容易にするために、ポイズニングされた異なるサンプルに異なるターゲットラベルを割り当てる。

表4は、データポイズニング攻撃に関する文献をまとめ、様々なモダリティと AI モデルタイプへの影響を示している。

データポイズニングのリスク緩和策には、通常、次のものが含まれる：学習データの厳格な

検証。潜在的に悪意のあるデータポイントをフラグ付けするための異常検知手法。およびポイズニングされた入力に対するモデルの耐性を向上させるために特別に設計された学習手順。

5.5. E：回避攻撃

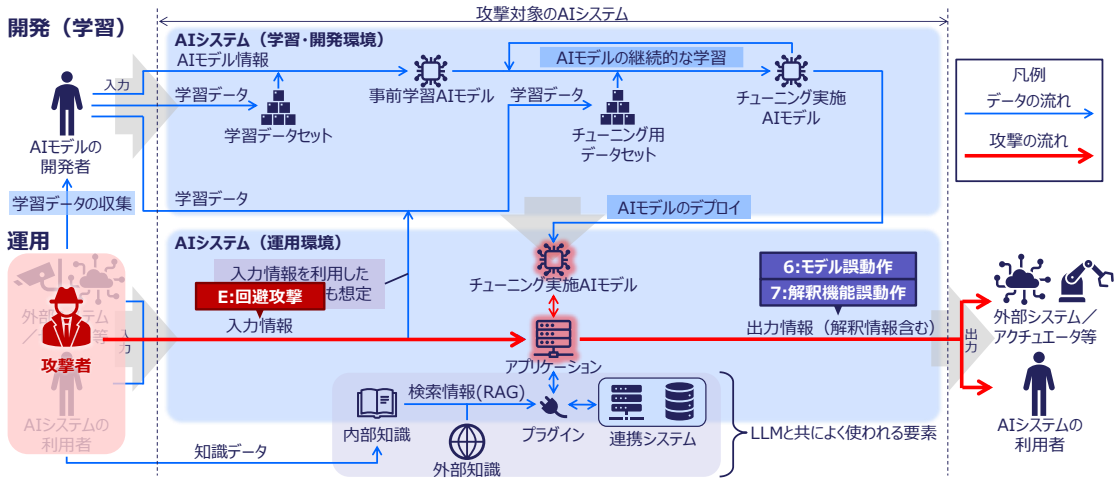

図 10　E：回避攻撃（概要）

回避攻撃とは、学習データを汚染したりモデル自体を直接改変したりすることなく、**敵対的サンプル (Adversarial Examples)** として知られる巧妙に細工された入力を提供することによって、運用段階で AI モデルの挙動を攻撃者が操作する脅威を表す（図 10 参照）。敵対的サンプルは、モデルを騙して不正な予測を行わせるように設計されており、正規のデータに巧妙な摂動を導入することによって生成される。

敵対的サンプルは、主に 2 つの条件下で生成され得る：

- **ホワイトボックス条件**：攻撃者がモデルの内部アーキテクチャとパラメータの完全な知識を持つ。
- **ブラックボックス条件**：攻撃者が内部詳細へのアクセスなしに、モデルの入力と出力のみを観測する。

どちらの条件においても、モデルの決定境界を悪用するために小さいながらも意図的な変更が入力に加えられ、モデルの出力に顕著な*逸脱*をもたらす。

この手法に関する議論の発端となった論文[30]の図 1 で、このような回避攻撃のデモンストレーションを示している。この例では、視覚的に知覚できない小さな摂動がパンダの画像に

加えられる。視覚的な違いは最小限であるにもかかわらず、この敵対的サンプルはモデルを首尾よく騙し、画像をテナガザルに誤分類させてしまう。回避攻撃が重大な*モデル誤動作*を誘発し得ることの実証となっている。

回避攻撃のもう一つの重大な結果である*解釈機能誤動作*は、別の論文[24]の図 1 で示されている。解釈機能誤動作では、モデルの説明メカニズムが特に攻撃の標的となり、説明可能性確保のための手法（例：ヒートマップ）によって意思決定プロセスに影響すると特定された領域が変更される。AI の動作に関する説明が操作・改竄されてしまうと、これらの視覚的解釈に依存して検証や意思決定を行うステークホルダーを、危険なほど誤った方向に導く可能性がある。

表 4 は、異なるモダリティと AI モデルタイプに対して、これらの影響を引き起こす回避攻撃に関する文献の概要をまとめている。

回避攻撃に対する一般的なリスク緩和戦略は、モデルの堅牢性を高めることと、評価・監視フレームワークを実装することを含む。注目すべきアプローチには以下が含まれる。

- 敵対的学習：敵対的サンプルを学習データセットに積極的に組み込むことによって、モデルの堅牢性を強化する。
- 堅牢性評価ツール：ART [61]、RobustBench [21]、CleverHans [70]といった専門のフレームワークを採用し、敵対的摂動に対するモデルの耐性を体系的にベンチマークする。
- 入力アクセス制御：AI システムが処理する入力の数と頻度を制限し、攻撃者がモデルの挙動を探る機会を減らす。
- 敵対的データ検出技術：高度な検出手法を活用し、敵対的に操作された入力を識別し影響を緩和する[3][53][96]。

解釈機能誤動作に焦点を当てた緩和手法もある。例えばβ平滑化は、勾配伝播をより滑らかなものにすることで、敵対的操作に対する堅牢性を高める[24]。

5.6. F：スポンジ攻撃

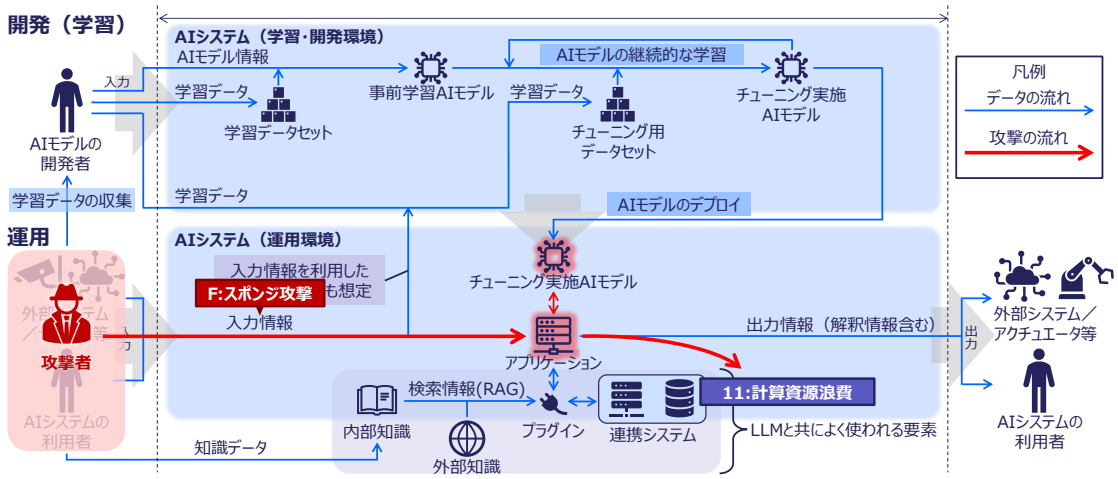

図 11　F：スポンジ攻撃（概要）

図 11 に示すスポンジ攻撃は、AI モデルの脆弱性を狙ったスポンジ・データ（Sponge Examples）と呼ばれる特殊な入力を与えることにより、推論時の計算負荷を増大させる。予測を誤らせるように設計された敵対的サンプルとは対照的に、リソースの枯渇、すなわち計算資源を浪費させることにスポンジ・データは焦点を当てており、エネルギー消費と応答遅延を大幅に増加させる。

表 4 にまとめられているように、スポンジ攻撃は画像やテキストを含む様々なモダリティで実証されている。先行研究では、これらの攻撃は必ずしも AI モデルの内部知識を必要とせず、公開 API との外部インタラクションでさえ、スポンジ攻撃のエクスプロイトを成功させるのに十分である可能性が示されている。

これらの攻撃の影響は、AI システムがリソース制約のある環境や遅延が致命的となるアプリケーションで運用されているシナリオで特に深刻である。反復的または持続的なスポンジ・データにさらされたシステムは、サービス可用性の低下、運用コストの増大、あるいは完全なサービス拒否（DoS）に見舞われる可能性がある。

リスク緩和戦略には、エネルギー消費や遅延が所定の制限を超えた場合に処理を終了させるカットオフ閾値を設定し、システム過負荷を防ぐエラーを発生させることが含まれる。また、より単純な処理に切り替えるなど、過度の計算負荷がかかった場合に作動するフォールバックメカニズムを導入することが、システムの安定性と可用性の維持に役立つ。

5.7. G：プロンプト窃盗攻撃

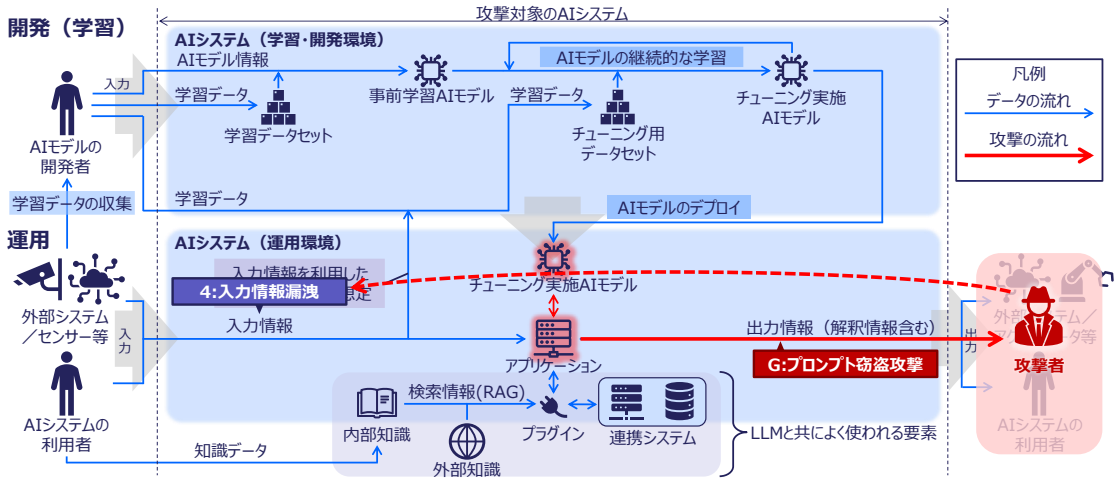

図 12　G：プロンプト窃盗攻撃（概要）

プロンプト窃盗攻撃は、AI によるコンテンツ生成の元となったテキストプロンプトを再構成する（図 12 参照）。入力を操作して生成モデルの挙動を侵害するプロンプトインジェクションとは対照的に、プロンプト窃盗は、モデルの出力を分析して元の入力をリバースエンジニアリングする。このような攻撃は、特に文章から画像を生成する拡散モデルを標的とし、慎重に設計されたプロンプトに組み込まれた知的財産にとって顕著な脅威となりえ、ひいてはこれらのプロンプトが商業的に取引される市場を混乱させうる（表 4 参照）。

最近の研究[79]は、実用的なプロンプト窃盗手法を提示している。この手法は、*被写体ジェネレータ*と*修飾子検出器*という、2 つの補完的な予測 AI コンポーネントと生成 AI コンポーネントを使用する。*被写体ジェネレータ*は画像からキャプションを生成する AI モデルとして実装され、与えられた画像に描かれた主要な被写体のテキスト表現を推論・生成する。*修飾子検出器*は、マルチラベル分類器として機能する予測 AI モデルとして実装され、視覚的詳細や芸術的スタイルを決定する細かなキーワードやスタイル修飾子を特定する。被写体と修飾子それぞれの予測結果を組み合わせることで、この手法は窃盗対象と酷似した詳細なプロンプトを効果的に再構成する。

プロンプト窃盗攻撃は生成された画像を盗むものではなく、攻撃者の手元には生成済みの画像があることが前提となる。これらの攻撃は、生成済み画像の根底にあるテキストプロンプトを標的とする。そのようなプロンプトは実用的かつ経済的な価値を有するからである。元のプロンプトへのアクセスにより、攻撃者は特定の芸術的スタイルの画像を複製したり、関連する創作活動に合わせてプロンプトを自由に調整したり、（訳注：本来であれば有償でそ

のプロンプトを購入することになる）プロンプト市場での金銭的な支払いや知的財産権の制限を回避したりすることが可能になる。

研究[79]では、プロンプト窃盗攻撃への対策も提案されている。具体的には、画像への敵対的摂動の組み込みを行うものである。ここで言う敵対的摂動とは、画像に加える微細で人間には知覚できない程度の変更であり、これにより攻撃者が元のプロンプトを正確に推論することを妨げる。しかし、高度な手法を用いる適応型の敵対者がこれらの防御を部分的に克服する可能性も指摘されており、堅牢かつ適応型のリスク緩和戦略が依然として必要であることが強調されている。

5.8. H：プロンプトインジェクション攻撃

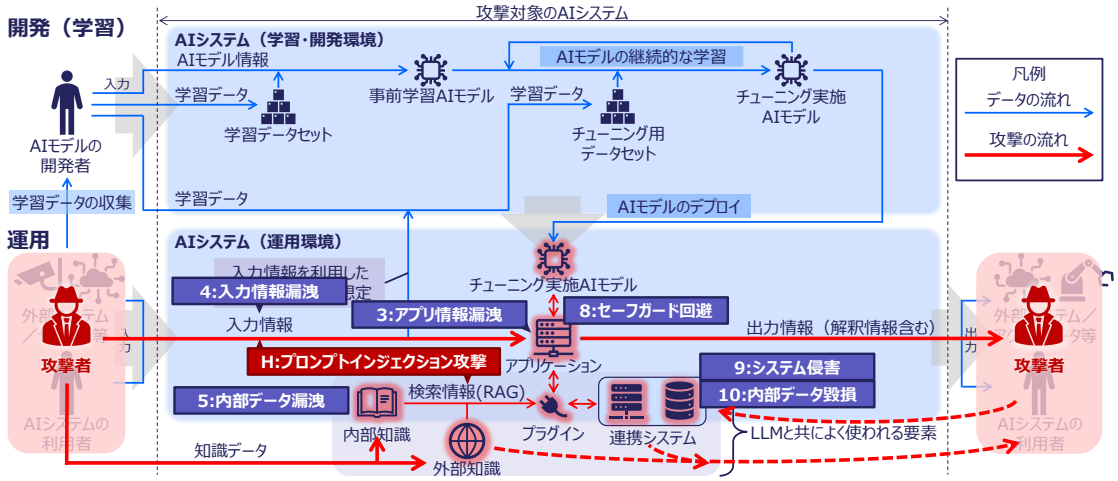

図 13　H：プロンプトインジェクション攻撃（概要）

プロンプトインジェクション攻撃は、AI システムへの入力プロンプト中に悪意のある指示を埋め込むことで、AI システムにおいて意図されていない、あるいは、有害な動作を実行させる（図 13）。これらの攻撃は 2 つのタイプに分類できる：

- **直接プロンプトインジェクション**：有害なコマンドが利用者の入力に直接含まれる。
- **間接プロンプトインジェクション**：攻撃者が、AI システムによってアクセスされる外部の情報源（例：データベース、API、またはプラグイン）を操作する。（訳注：これにより、正規の利用者の入力するプロンプトに悪意ある指示が混入するよう仕向ける）

プロンプトインジェクション攻撃の影響は広範にわたり、セキュリティのセーフガード回避、様々な種類の情報漏洩、不正なコマンドの実行などがある（表 4 参照）。

プロンプトインジェクション攻撃に対する様々な緩和戦略が提案されている。AI モデルの

外側で実装される外部対策としては、悪意のある入力や危険な出力の検出とフィルタリング[35]、システムプロンプトによるモデルの挙動制御[95]、プロンプト内で信頼できるデータと信頼できないデータを分離すること[36]などが挙げられる。更に、防御メカニズムを AI モデルに直接統合することも普及している。そのような手法には、悪意のあるプロンプトに対するモデルの耐性を高めるための敵対的学習[27]やその他の特定の学習[105]が含まれる。

5.9. I：コードインジェクション攻撃

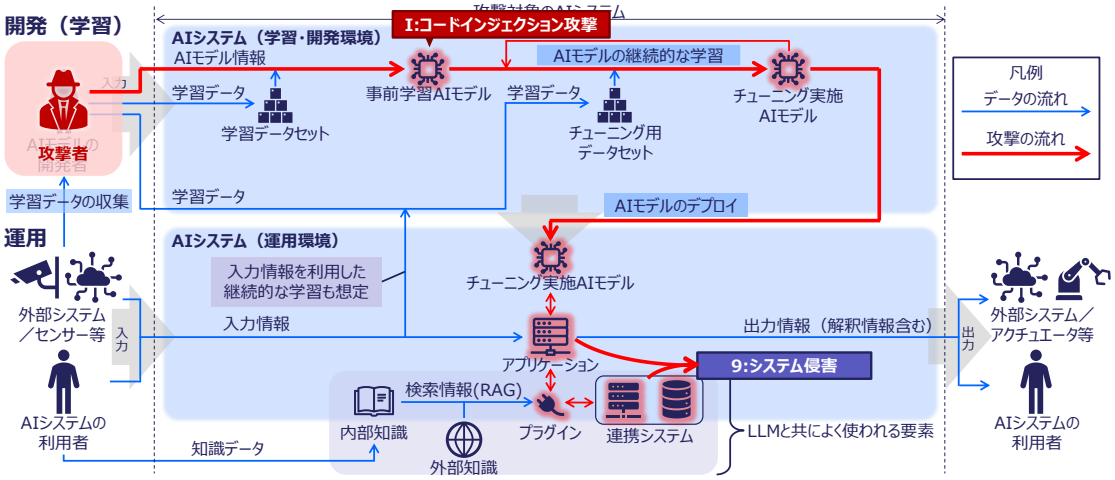

図 14　I：コードインジェクション攻撃（概要）

OWASP[104]によれば、**コードインジェクション**とは、「アプリケーションによって解釈・実行されるコードを注入することからなる攻撃タイプの総称」である。本稿では、AI モデルへのコードインジェクションに焦点を当てる。即ち、悪意のある実行可能コードがモデル内に埋め込まれ、モデルのロード時に実行される攻撃である。推論時にモデルの入力を操作するプロンプトインジェクションとは対照的に、コードインジェクションは悪意のあるコードをモデルファイル自体に直接埋め込む。

コードインジェクションが及ぼす主な影響はシステム侵害であり、これにより攻撃者は不正なコマンドを侵害先のシステム上で実行できるようになる。他の攻撃とは異なり、データのモダリティとは無関係に、コードインジェクションは予測 AI モデルおよび生成 AI モデルの両方にリスクをもたらす（表 4）。

コードインジェクションは、Python の `pickle` のような、コード実行をサポートするよう意図的に設計されているデータフォーマットを悪用する。攻撃者は、そのようなフォーマッ

ト（例：pickle、TensorFlow の SavedModel、Keras の Lambda レイヤー）を狙い、悪意のあるペイロードをモデルに埋め込む。これらの悪意のあるペイロードを含んだモデルデータは、通常、攻撃者によってパブリックリポジトリにアップロードされる。適切な検証なしにこれらのモデルデータがダウンロードおよび実行されると、注入されたコードも自動的に実行される（図 14 参照）。これらのフォーマットにおける実行可能コードの埋め込みは、それを許すフォーマット自体は脆弱性に当たるものではなく、正当に意図された仕様が攻撃者によって悪用されていることに留意すべきである。

コードインジェクションのリスク緩和策には、GGUF、ONNX、Safetensors といった、本質的に安全なフォーマットの使用が含まれる。これらのフォーマットでは実行可能コードの埋め込みがサポートされておらず、モデルの重み（とアーキテクチャ）のみがデータに含まれる。その他の対策には、モデルの出所の検証、サンドボックス環境の採用、悪意のあるコードを検出するための分析ツールの活用[103]などがある。

5.10. J：ファインチューニング攻撃

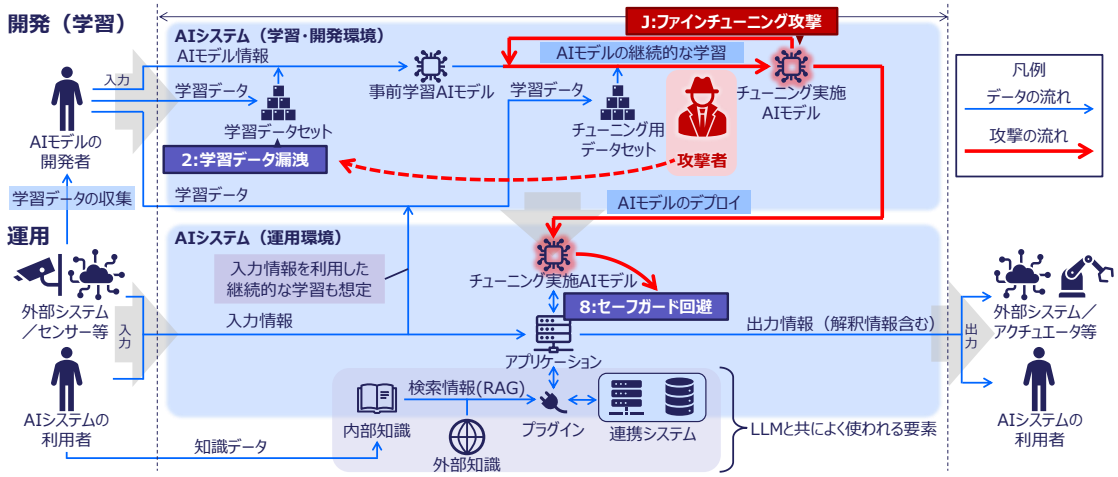

図 15　J：ファインチューニング攻撃（概要）

敵対的なファインチューニング攻撃では、AI モデルのファインチューニングプロセスを悪用し、意図されたセーフティ・アラインメントを侵害する。ファインチューニングは事前学習済みの LLM を専門タスクに適応させる際によく実施されるが、このカテゴリの攻撃はこのファインチューニング工程に焦点を当てている点に特徴がある。これらの攻撃は本来のセーフガードを回避するものであり、結果として、有害な挙動を誘発したり、モデルの安全性を損なったり、以前に学習したデータから機密情報を再構成したりするように構成されている（図 15 参照）。ファインチューニング攻撃の基本メカニズムには、モデルのセーフティ・

アラインメントを効果的に覆したり、以前に抑制されていた機密情報を再アクティブ化したりする、特殊なファインチューニング手順を実行することが伴っている。

最近の研究[18]では、慎重に選ばれた限定的なサンプルセットでファインチューニングを行うと、プライバシー漏洩の危険性が大幅に増す可能性のあることが実証されている。同様に、別の研究[100]では、オープンアクセスの LLM においてアラインメントを容易に覆し、それによってモデルが有害または非倫理的な応答を生成するようになることが実証されている（表 4 参照）。

ファインチューニング攻撃に対するリスク緩和戦略には、いくつかの対策が含まれる。例えば、ファインチューニングインターフェースへのアクセス制御の実装や、モデルの挙動の継続的な監視が該当する。

5.11. K：ロウハンマー攻撃

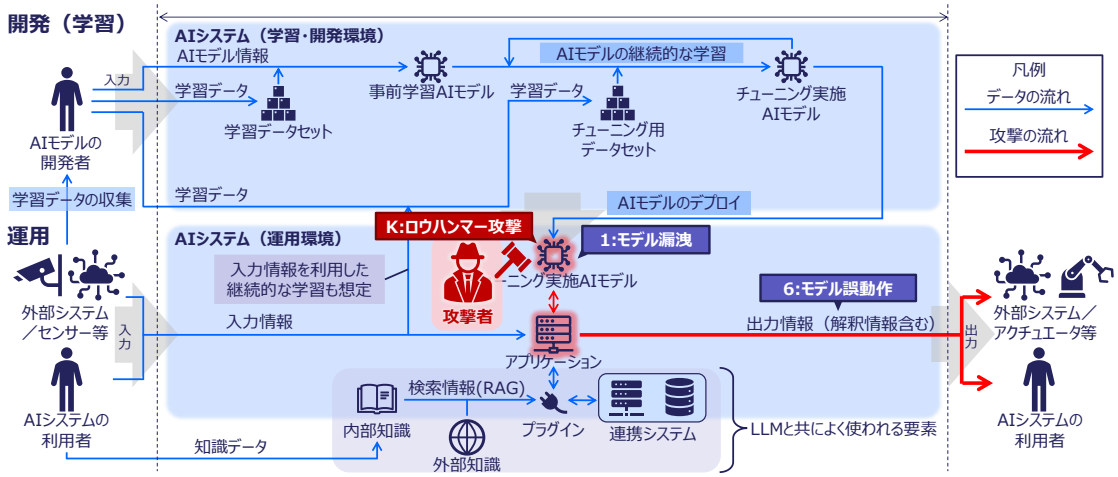

図 16　K：ロウハンマー攻撃（概要）

ロウハンマー攻撃は、DRAM モジュールのハードウェアレベルの脆弱性を悪用し、メモリアクセスを繰り返すことで不正なビット反転を引き起こす。他の多くの攻撃とは対照的に、ロウハンマーは AI モデルのパラメータを格納する物理メモリセルを標的とし、予測型と生成型の両方の AI モデルの侵害をもたらす。

ロウハンマーの基本メカニズムは、特定のメモリ行（Memory Row）への反復的なアクセスであり、これが電気的干渉を誘発する。結果として、隣接するメモリ行での意図しないビット反転を引き起こす。これらの意図しないビット反転は、実行時に重要なモデルパラメータや実行コードを破壊する可能性がある（図 16）。

例えば、DeepSteal[74]のような攻撃では、ロウハンマーによって誘発されたビット反転を悪用して、ニューラルネットワークモデルから部分的な重みを遠隔で抽出してみせ、最終的に、被害モデルの不正な複製に成功している。同様に、FrameFlip[46]攻撃は、内部構造の詳細な知識なしに実行時コードに欠陥を注入することで、ディープニューラルネットワークの推論性能を広く低下させられることを実証した。

ロウハンマーの影響には、モデル誤動作またはモデル漏洩も含まれる（表 4 参照）。最近の研究では、たった 1 回のビット反転でさえ、ニューラルネットワークの性能を大幅に低下させたり変化させたりする可能性が実証されており、ハードウェアレベルの摂動に対する AI モデルの推論の感度の高さが浮き彫りになっている[46]。

ロウハンマー攻撃に対するリスク緩和戦略は、ハードウェアからソフトウェアレベルにまで及ぶ。ハードウェア側では、ターゲット行リフレッシュのような、より強力なエラー訂正を備えた改良型 DRAM 設計などの解決策がある。ソフトウェアベースの対策は、メモリアロケーションのランダム化など、堅牢なメモリ管理プラクティスに焦点を当てている。重要な重みを隠すためのモデルニューロンの戦略的な「リワイヤリング」など、最近提案された技術も、トランスフォーマーベースのモデルのビット反転攻撃に対する耐性を高める上で有効であることが示されている[60]。

6. まとめ

本稿では、AI システムを標的とする既知の敵対的攻撃の概要を提供し、その根底にあるメカニズム、影響、および攻撃を成立させるための諸条件について詳述した。これらの攻撃をその技術的方法論に従って分類することにより、特に AI セキュリティの専門知識を持たない可能性のある研究者、開発者、セキュリティ実務者、政策立案者にとって、理解と実用的な応用の両方を促進する、明確で直感的な分類体系を提示している。

我々の分析は、学習からデプロイに至る複数の段階にわたる AI システム固有の脆弱性を浮き彫りにすることによって、AI システムの保護が必要不可欠であることを強調した。我々は、個々の攻撃と、サイバーセキュリティにおける中核コンポーネントである機密性・完全性・可用性へ攻撃が及ぼす潜在的な影響との関連性を体系的に図解した。本稿全体で提供される直感的で視覚的なマッピングは、実務者が自身の AI デプロイに関連する脆弱性を迅速に特定し、的を絞った緩和戦略を策定する助けとなることを意図している。

更に、我々は NIST AI 100-2e2025 や MITRE ATLAS™を含む既存の文献を広範にレビューした上で、基礎的研究の成果を最先端の取り組みと統合し、先行研究の限界を特定した。追

加の洞察を提供することによって、我々の成果は既存のフレームワークを補完するだけでなく、レッドチーム演習のような実際のセキュリティ管理シナリオに適用可能な実践的ガイダンスも提供するものである。

AI 技術が急速に進歩し続けるにつれて、新たな脅威も必然的に出現し、脅威モデリングのフレームワークと防御戦略を継続的に更新し拡張する必要が伴うものと予期される。したがって、将来の研究では、新たに出現する AI の脆弱性の更なる探求、防御法の改良、そして進化する AI セキュリティの課題にあらゆるレベルのステークホルダーが積極的に対処できる備えを確実にするための、知識を普及する努力の拡大に焦点を当てるべきである。

突き詰めれば、AI セキュリティの向上は単なる技術的責務ではなく、より広範な社会的責任である。それには、企業統治、堅牢な技術的防御、包括的な政策決定、そして進行中の国際協力を含む統合的アプローチが必要である。本稿は、これらの課題に対処し、より安全で、信頼性が高く、責任ある AI システムを促進するために必要な基礎知識に貢献するものと信じる。

謝辞

著者らは J-AISI と IPA の同僚各位の継続的な支援に感謝申し上げる。本研究は、秋間建人氏、北村弘氏、栗原悠磨氏、中里克久氏、松岡光氏から提供された貴重な助言なしには成り立たなかった。本稿のドラフトの一部を効率的に執筆するために、生成 AI ツール、具体的には OpenAI ChatGPT と Google Gemini を利用した。著者らは本稿のすべての内容をレビューし、検証している。

参考文献

[1] Monika Adamczyk, Nineta Polemi, Isabel Praça, and Konstantinos Moulinos. 2023. A multilayer framework for good cybersecurity practices for AI: security and resilience for smart health services and infrastructures. Technical Report. European Union Agency for Cybersecurity. doi:10.2824/588830

[2] Salem Ahmed, Zhang Yang, Humbert Mathias, Berrang Pascal, Fritz Mario, and Backes Michael. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In Proceedings 2019 Network and Distributed System Security Symposium (San Diego, CA, USA) (NDSS 2019). Internet Society, Reston, VA, USA. doi:10.14722/ndss.2019.23119

[3] Ahmed Aldahdooh, Wassim Hamidouche, Sid Ahmed Fezza, and Olivier Déforges. 2022. Adversarial Example Detection for DNN Models: A Review and Experimental Comparison. Artificial Intelligence Review 55, 6 (Aug. 2022), 4403–4462. doi:10.1007/s10462-021-10125-w

[4] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. 2015. Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers. International Journal of Security and Networks 10, 3 (Sept. 2015), 137–150. doi:10.1504/IJSN.2015.071829

[5] Borja Balle, Giovanni Cherubin, and Jamie Hayes. 2022. Reconstructing Training Data with Informed Adversaries. In 2022 IEEE Symposium on Security and Privacy (SP) (San Francisco, CA, USA). IEEE Computer Society, Los Alamitos, CA, USA, 1138–1156. doi:10.1109/SP46214.2022.9833677

[6] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. 2022. Bad Characters: Imperceptible NLP Attacks. In 2022 IEEE Symposium on Security and Privacy (SP) (San Francisco, CA, USA). IEEE Computer Society, Los Alamitos, CA, USA, 1987–2004. doi:10.1109/SP46214.2022.9833641

[7] Gon Buzaglo, Niv Haim, Gilad Yehudai, Gal Vardi, and Michal Irani. 2023. Reconstructing Training Data from Multiclass Neural Networks. arXiv:2305.03350 [cs.LG]

[8] Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. 2021. IPGuard: Protecting Intellectual Property of Deep Neural Networks via Fingerprinting the Classification Boundary. In Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security (Virtual Event, Hong Kong) (ASIA CCS '21). Association for Computing Machinery, New York, NY, USA, 14–25. doi:10.1145/3433210.3437526

[9] Nicholas Carlini. 2021. Poisoning the Unlabeled Dataset of Semi-Supervised Learning. In 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, Berkeley, CA, USA, 1577–1592. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-poisoning

[10] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting Training Data from Diffusion Models. In 32nd USENIX Security Symposium (USENIX Security 23) (Anaheim, CA). USENIX Association, Berkeley, CA, USA, 5253–5270. https://www.usenix.org/conference/usenixsecurity23/presentation/carlini

[11] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In 28th USENIX Security Symposium (USENIX Security 19) (Santa Clara, CA, USA). USENIX Association, Berkeley, CA, USA, 267–284. https://www.usenix.org/conference/usenixsecurity19/presentation/carlini

[12] Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. 2023. Are aligned neural networks adversarially aligned?. In Advances in Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '23, Vol. 36), A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.). Curran Associates, Inc., Red Hook, NY, USA, 61478– 61500. https://proceedings.neurips.cc/paper_files/paper/2023/hash/c1f0b856a35986348ab3414177266f75-Abstract-Conference.html

[13] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In 30th USENIX Security Symposium (USENIX Security 21). USENIX

Association, Berkeley, CA, USA, 2633–2650.
https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

[14] National Cyber Security Centre. 2024. Machine learning principles.
https://www.ncsc.gov.uk/collection/machine-learning-principles

[15] National Cyber Security Centre, Cybersecurity and Infrastructure Security Agency, National Security Agency, Federal Bureau of Investigation, Australian Signals Directorate's Australian Cyber Security Centre, Canadian Centre for Cyber Security, New Zealand National Cyber Security Centre, Chile's Government CSIRT, National Cyber and Information Security Agency of the Czech Republic, Information System Authority of Estonia, National Cyber Security Centre of Estonia, French Cybersecurity Agency, Germany's Federal Office for Information Security, Israeli National Cyber Directorate, Italian National Cybersecurity Agency, Japan's National center of Incident readiness and Strategy for Cybersecurity, Japan's Secretariat of Science, Technology and Innovation Policy, Cabinet Office, Nigeria's National Information Technology Development Agency, Norwegian National Cyber Security Centre, Poland Ministry of Digital Affairs, Poland's NASK National Research Institute, Republic of Korea National Intelligence Service, and Cyber Security Agency of Singapore. 2023. Guidelines for secure AI system development.
https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

[16] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2024. Jailbreaking Black Box Large Language Models in Twenty Queries. arXiv:2310.08419 [cs.LG]

[17] Huaming Chen and M. Ali Babar. 2024. Security for Machine Learning-based Software Systems: A Survey of Threats, Practices, and Challenges. ACM Comput. Surv. 56, 6, Article 151 (Feb. 2024), 38 pages. doi:10.1145/3638531

[18] Xiaoyi Chen, Siyuan Tang, Rui Zhu, Shijun Yan, Lei Jin, Zihao Wang, Liya Su, Zhikun Zhang, XiaoFeng Wang, and Haixu Tang. 2024. The Janus Interface: How Fine-Tuning in Large Language Models Amplifies the Privacy Risks. In Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security (Salt Lake City, UT, USA) (CCS '24). Association for Computing Machinery, New York, NY, USA, 1285–1299. doi:10.1145/3658644.3690325

[19] Antonio Emanuele Cinà, Ambra Demontis, Battista Biggio, Fabio Roli, and Marcello Pelillo. 2023. Energy-Latency Attacks via Sponge Poisoning. arXiv:2203.08147 [cs.CR]

[20] Jacson Rodrigues Correia-Silva, Rodrigo F. Berriel, Claudine Badue, Alberto F. de Souza, and Thiago Oliveira-Santos. 2018. Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data. In 2018 International Joint Conference on Neural Networks (IJCNN) (Rio de Janeiro, Brazil). IEEE, New York, NY, USA, 1–8. doi:10.1109/IJCNN.2018.8489592

[21] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2021. RobustBench: a standardized adversarial robustness benchmark. In Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks, J. Vanschoren and S. Yeung (Eds.). Neural Information Processing Systems Founda- tion, San Diego, CA, USA. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/a3c65c2974270fd093ee8a9bf8ae7d0b-Abstract-round2.html

[22] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2025. Security and Privacy Challenges of Large Language Models: A Survey. ACM Comput. Surv. 57, 6, Article 152 (Feb. 2025), 39 pages. doi:10.1145/3712001

[23] Digital Architecture Research Center, Cyber Physical Security Research Center, and Artificial Intelligence Research Center. 2024. Machine Learning Quality Management Guidline (Revision 4.2.0). Technical Report. National Institute of Advanced Industrial Science and Technology (AIST). https://www.digiarc.aist.go.jp/publication/aiqm/guideline-rev4.html in Japanese.

[24] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations can be manipulated and geometry is to blame. In Advances in Neural Information Processing Systems (Vancouver, Canada), H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., Red Hook, NY, USA. https://proceedings.neurips.cc/paper/2019/hash/bb836c01cdc9120a9c984c525e4b1a

4a-Abstract.html

[25] Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. 2024. Attacks, Defenses and Evaluations for LLM Conversation Safety: A Survey. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (Mexico City, Mexico), Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 6734–6747. doi:10.18653/v1/2024.naacl-long.375

[26] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Melbourne, Australia), Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 31–36. doi:10.18653/v1/P18-2006

[27] Aidan Ewart, Abhay Sheshadri, Phillip Huang Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. 2024. Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs. (2024). https://neurips.cc/virtual/2024/103310 Workshop on Socially Responsible Language Modelling Research (SoLaR) at NeurIPS.

[28] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (Denver, Colorado, USA) (CCS '15). Association for Computing Machinery, New York, NY, SA, 1322–1333. doi:10.1145/2810103.2813677

[29] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. 2018. Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (Toronto, Canada) (CCS '18). Association for Computing Machinery, New York, NY, USA, 619–633. doi:10.1145/3243734.3243834

[30] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and

Harnessing Adversarial Examples. arXiv:1412.6572 [stat.ML] 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

[31] Justin Grana. 2020. Perturbing Inputs to Prevent Model Stealing. In 2020 IEEE Conference on Communications and Network Security (CNS) (Avignon, France). IEEE, New York, NY, USA, 1–9. doi:10.1109/CNS48642.2020.9162336

[32] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (Copenhagen, Denmark) (AISec '23). Association for Computing Machinery, New York, NY, USA, 79–90. doi:10.1145/3605764.3623985

[33] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based Adversarial Attacks against Text Transformers. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (Punta Cana, Dominican Republic), Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 5747–5757. doi:10.18653/v1/2021.emnlp-main.464

[34] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. 2022. Reconstructing Training Data from Trained Neural Networks. In Proceedings of the 36th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 1665, 14 pages. https://proceedings.neurips.cc/paper_files/paper/2022/file/906927370cbeb537781100623cca6fa6-Paper-Conference.pdf

[35] Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. WildGuard: Open One-stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs. In The Thirty-eight Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks (Vancouver, BC, Canada), A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran

Associates, Inc., Red Hook, NY, USA, 8093–8131. https://proceedings.neurips.cc/paper_files/paper/2024/hash/0f69b4b96a46f284b726fbd70f74fb3b-Abstract-Datasets_and_Benchmarks_Track.html

[36] Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kıcıman. 2024. Defending Against Indirect Prompt Injection Attacks With Spotlighting. In Conference on Applied Machine Learning for Informa- tion Security (CAMLIS 2024) (Arlington, VA, USA), Rachel Allen, Sagar Samtani, Edward Raff, and Ethan Rudd (Eds.), Vol. 3920. CEUR-WS.org, Aachen, Germany, 48–62. https://ceur-ws.org/Vol-3920/paper03.pdf

[37] Yupeng Hu, Wenxin Kuang, Zheng Qin, Kenli Li, Jiliang Zhang, Yansong Gao, Wenjia Li, and Keqin Li. 2021. Artificial Intelligence Security: Threats and Countermeasures. ACM Comput. Surv. 55, 1, Article 20 (Nov. 2021), 36 pages. doi:10.1145/3487890

[38] Japan AI Safety Institute (J-AISI). 2024. Guide to Red Teaming Methodology on AI Safety. https://aisi.go.jp/output/output_framework/guide_to_red_teaming_methodology_on_ai_safety/

[39] Japan AI Safety Institute (J-AISI). 2024. Overview of the Japan AI Safety Institute (J-AISI). https://aisi.go.jp/about/

[40] Japan AI Safety Institute (J-AISI). 2025. Known Attacks and Their Impacts on AI Systems. https://aisi.go.jp/output/output_security/known_attacks_and_impacts/

[41] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. 2018. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. In 2018 IEEE Symposium on Security and Privacy (SP) (San Francisco, CA, USA). IEEE, New York, NY, USA, 19–35. doi:10.1109/SP.2018.00057

[42] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N. Asokan. 2019. PRADA: Protecting Against DNN Model Stealing Attacks. In 2019 IEEE European Symposium on Security and Privacy (EuroS&P) (Stockholm, Sweden). IEEE Computer Society, Los Alamitos, CA, USA, 512–527. doi:10.1109/EuroSP.2019.00044

[43] Yusuke Kawamoto, Kazumasa Miyake, Koichi Konishi, and Yutaka Oiwa. 2023.

Threats, Vulnerabilities, and Controls of Machine Learning Based Systems: A Survey and Taxonomy. arXiv:2301.07474 [cs.CR]

[44] Zixiao Kong, Jingfeng Xue, Yong Wang, Lu Huang, Zequn Niu, and Feng Li. 2021. A Survey on Adversarial Attack in the Age of Artificial Intelligence. Wireless Communications and Mobile Computing 2021, 1 (2021), 4907754. doi:10.1155/2021/4907754

[45] Peter Kyle. 2025. Tackling AI security risks to unleash growth and deliver Plan for Change. https://www.gov.uk/government/news/tackling-ai-security-risks-to-unleash-growth-and-deliver-plan-for-change Secretary of State for Science, Innovation and Technology, UK Government.

[46] Shaofeng Li, Xinyu Wang, Minhui Xue, Haojin Zhu, Zhi Zhang, Yansong Gao, Wen Wu, and Xuemin (Sherman) Shen. 2024. Yes, One-Bit-Flip Matters! Universal DNN Model Inference Depletion with Runtime Code Fault Injection. In 33rd USENIX Security Symposium (USENIX Security 24) (Philadelphia, PA, USA). USENIX Association, Berkeley, CA, USA, 1315–1330. https://www.usenix.org/conference/usenixsecurity24/presentation/li-shaofeng

[47] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. 2024. Backdoor Learning: A Survey. IEEE Transactions on Neural Networks and Learning Systems 35, 1 (2024), 5–22. doi:10.1109/TNNLS.2022.3182979 arXiv:2007.08745 [cs.CR]

[48] Zongjie Li, Chaozheng Wang, Shuai Wang, and Cuiyun Gao. 2023. Protecting Intellectual Property of Large Language Model-Based Code Generation APIs via Watermarks. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (Copenhagen, Denmark) (CCS '23). Association for Computing Machinery, New York, NY, USA, 2336–2350. doi:10.1145/3576915.3623120

[49] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2024. Prompt Injection attack against LLM-integrated Applications. arXiv:2306.05499 [cs.CR]

[50] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. 2022. ML-Doctor: Holistic Risk

Assessment of Inference Attacks Against Machine Learning Models. In 31st USENIX Security Symposium (USENIX Security 22) (Boston, MA, USA). USENIX Association, Berkeley, CA, USA, 4525–4542. https://www.usenix.org/conference/usenixsecurity22/presentation/liu-yugeng

[51] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen. 2018. Understanding Membership Inferences on Well-Generalized Learning Models. arXiv:1802.04889 [cs.CR]

[52] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing Leakage of Personally Identifiable Information in Language Models. In 2023 IEEE Symposium on Security and Privacy (SP) (San Francisco, CA, USA). IEEE, New York, NY, USA, 346–363. doi:10.1109/SP46215.2023.10179300

[53] Shiqing Ma, Yingqi Liu, Guanhong Tao, Wen-Chuan Lee, and Xiangyu Zhang. 2019. NIC: Detecting Adversarial Samples with Neural Network Invariant Checking. In Proceedings 2019 Network and Distributed System Security Symposium (San Diego, CA, USA) (NDSS 2019). Internet Society, Reston, VA, USA. doi:10.14722/ndss.2019.23415

[54] Saeed Mahloujifar, Esha Ghosh, and Melissa Chase. 2022. Property Inference from Poisoning. In 2022 IEEE Symposium on Security and Privacy (SP) (San Francisco, CA, USA). IEEE, New York, NY, USA, 1120–1137. doi:10.1109/SP46214.2022.9833623

[55] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of Attacks: Jailbreaking Black-Box LLMs Automatically. arXiv:2312.02119 [cs.LG]

[56] MITRE. 2013. MITRE ATT&CK®. https://attack.mitre.org/

[57] MITRE. 2021. MITRE ATLAS™. https://atlas.mitre.org/

[58] Sasi Kumar Murakonda and Reza Shokri. 2020. ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning. (2020). arXiv:2007.09339 [cs.CR] Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs).

[59] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In 2019 IEEE Symposium on Security and Privacy (SP) (San Francisco, CA, USA). IEEE, New York, NY, USA, 739–753. doi:10.1109/SP.2019.00065

[60] Najmeh Nazari, Hosein Mohammadi Makrani, Chongzhou Fang, Hossein Sayadi, Setareh Rafatirad, Khaled N. Khasawneh, and Houman Homayoun. 2024. Forget and Rewire: Enhancing the Resilience of Transformer-based Models against Bit-Flip Attacks. In 33rd USENIX Security Symposium (USENIX Security 24) (Philadelphia, PA, USA). USENIX Association, Berkeley, CA, USA, 1349–1366. https://www.usenix.org/conference/usenixsecurity24/presentation/nazari

[61] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian M. Molloy, and Ben Edwards. 2019. Adversarial Robustness Toolbox v1.0.0. arXiv:1807.01069 [cs.LG]

[62] Stavros Ntalampiras, Corina Pascu, Marco Barros Lourenco, Gianluca Misuraca, and Pierre Rossel. 2023. Artificial intelligence and cybersecurity research. Technical Report. European Union Agency for Cybersecurity. doi:10.2824/808362

[63] U.S. Department of Commerce. 2025. Statement from U.S. Secretary of Commerce Howard Lutnick on Trans- forming the U.S. AI Safety Institute into the Pro-Innovation, Pro-Science U.S. Center for AI Standards and Innova- tion. https://www.commerce.gov/news/press-releases/2025/06/statement-us-secretary-commerce-howard-lutnick-transforming-us-ai

[64] National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. doi:10.6028/NIST.AI.100-1

[65] National Institute of Standards and Technology. 2023. NIST AI RMF Playbook. https://airc.nist.gov/airmf-resources/playbook/

[66] The Leaders of the Group of Seven (G7). 2023. Hiroshima AI Process. https://www.soumu.go.jp/hiroshimaaiprocess/en/

[67] Seong Joon Oh, Bernt Schiele, and Mario Fritz. 2019. Towards Reverse-Engineering

Black-Box Neural Networks. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (Eds.). Lecture Notes in Computer Science, Vol. 11700. Springer International Publishing, Cham, 121–144. doi:10.1007/978-3-030-28954-6_7

[68] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2019. Knockoff Nets: Stealing Functionality of Black-Box Models. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Long Beach, CA, USA). IEEE, New York, NY, USA, 4949–4958. doi:10.1109/CVPR.2019.00509

[69] Ayodeji Oseni, Nour Moustafa, Helge Janicke, Peng Liu, Zahir Tari, and Athanasios Vasilakos. 2021. Security and Privacy for Artificial Intelligence: Opportunities and Challenges. arXiv:2102.04661 [cs.CR]

[70] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, Rujun Long, and Patrick McDaniel. 2018. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. arXiv:1610.00768 [cs.LG]

[71] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. Practical Black-Box Attacks against Machine Learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (Abu Dhabi, United Arab Emirates) (ASIA CCS '17). Association for Computing Machinery, New York, NY, USA, 506–519. doi:10.1145/3052973.3053009

[72] Rodrigo Pedro, Miguel E. Coimbra, Daniel Castro, Paulo Carreira, and Nuno Santos. 2025. Prompt-to-SQL Injections in LLM-Integrated Web Applications: Risks and Defenses. In 2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE) (Ottawa, ON, Canada). IEEE Computer Society, Los Alamitos, CA, USA, 76–88. doi:10.1109/ICSE55347.2025.00007

[73] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Abu Dhabi, United Arab Emirates), Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 3419–3448. doi:10.18653/v1/2022.emnlp-main.225

[74] Adnan Siraj Rakin, Md Hafizul Islam Chowdhuryy, Fan Yao, and Deliang Fan. 2022. DeepSteal: Advanced Model Extractions Leveraging Efficient Weight Stealing in Memories. In 2022 IEEE Symposium on Security and Privacy (SP) (San Francisco, CA, USA). IEEE, New York, NY, USA, 1157–1174. doi:10.1109/SP46214.2022.9833743

[75] Roman Samoilenko. 2023. New prompt injection attack on ChatGPT web version. Reckless copy-pasting may lead to serious privacy issues in your chat. Netograph. Retrieved April 25, 2025 from https://kajojify.github.io/articles/1_chatgpt_attack.pdf

[76] Oscar Schwartz. 2019. In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation. IEEE Spectrum Nov. (2019). https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation Updated: Jan. 2024.

[77] Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks. arXiv:2310.10844 [cs.CL]

[78] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security (Salt Lake City, UT, USA) (CCS '24). Association for Computing Machinery, New York, NY, USA, 1671–1685. doi:10.1145/3658644.3670388

[79] Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. 2024. Prompt Stealing Attacks Against Text-to-Image Generation Models. In 33rd USENIX Security

Symposium (USENIX Security 24) (Philadelphia, PA, USA). USENIX Association, Berkeley, CA, USA, 5823–5840. https://www.usenix.org/conference/usenixsecurity24/presentation/shen-xinyue

[80] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In 2017 IEEE Symposium on Security and Privacy (SP) (San Jose, CA, USA). IEEE, New York, NY, USA, 3–18. doi:10.1109/SP.2017.41

[81] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. 2021. Sponge Examples: Energy-Latency Attacks on Neural Networks. In 2021 IEEE European Symposium on Security and Privacy (EuroS&P) (Vienna, Austria). IEEE, New York, NY, USA, 212–231. doi:10.1109/EuroSP51992.2021.00024

[82] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (New York, NY, USA) (AIES '20). Association for Computing Machinery, New York, NY, USA, 180–186. doi:10.1145/3375627.3375830

[83] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. 2017. Machine Learning Models that Remember Too Much. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (Dallas, Texas, USA) (CCS '17). Association for Computing Machinery, New York, NY, USA, 587–601. doi:10.1145/3133956.3134077

[84] Congzheng Song and Vitaly Shmatikov. 2020. Overlearning Reveals Sensitive Attributes. arXiv:1905.11742 [cs.LG] https://openreview.net/forum?id=SJeNz04tDS 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.

[85] Hui Sun, Tianqing Zhu, Zhiqiu Zhang, Dawei Jin, Ping Xiong, and Wanlei Zhou. 2023. Adversarial Attacks Against Deep Generative Models on Data: A Survey. IEEE Transactions on Knowledge & Data Engineering 35, 04 (April 2023), 3367–3388. doi:10.1109/TKDE.2021.3130903

[86] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian

J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. arXiv:1312.6199 [cs.CV] 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings.

[87] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. Stealing Machine Learning Models via Prediction APIs. In 25th USENIX Security Symposium (USENIX Security 16) (Austin, TX, USA). USENIX Association, Berkeley, CA, USA, 601–618. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer

[88] Andrew van der Stock, Brian Glas, Neil Smithline, and Torsten Gigler. 2021. OWASP Top 10 (Version 2021). https://owasp.org/Top10/

[89] Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson. 2024. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. Technical Report NIST Artifcial Intelligence (AI) Report, NIST Trustworthy and Responsible AI NIST AI 100-2e2023. National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.AI.100-2e2023

[90] Apostol Vassilev, Alina Oprea, Alie Fordyce, Hyrum Anderson, Xander Davies, and Maia Hamin. 2025. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. Technical Report NIST Trustworthy and Responsible AI, NIST AI 100-2e2025. National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.AI.100-2e2025

[91] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Hong Kong, China), Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 2153–2162. doi:10.18653/v1/D19-1221

[92] Binghui Wang and Neil Zhenqiang Gong. 2018. Stealing Hyperparameters in Machine Learning. In 2018 IEEE Symposium on Security and Privacy (SP) (San Francisco, CA,

USA). IEEE, New York, NY, USA, 36–52. doi:10.1109/SP.2018.00038

[93] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How Does LLM Safety Training Fail?. In Advances in Neural Information Processing Systems (New Orleans, LA, USA), A. Oh, T. Naumann, A. Glober- son, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., Red Hook, NY, USA, 80079–80110. https://proceedings.neurips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html

[94] Steve Wilson and Ads Dawson. 2024. OWASP Top 10 for Large Language Model Applications (Version 2025). https://owasp.org/www-project-top-10-for-large-language-model-applications/

[95] Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023. Defending chatGPT against Jailbreak Attack via Self-Reminders. Nature Machine Intelligence 5, 12 (2023), 1486–1496. doi:10.1038/s42256-023-00765-8

[96] Weilin Xu, David Evans, and Yanjun Qi. 2018. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In Proceedings 2018 Network and Distributed System Security Symposium (San Diego, CA, USA) (NDSS 2018). Internet Society, Reston, VA, USA. doi:10.14722/ndss.2018.23198

[97] Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2025. On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review. High-Confidence Computing 5, 2 (2025), 100300. doi:10.1016/j.hcc.2025.100300

[98] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. High-Confidence Computing 4, 2 (2024), 100211. doi:10.1016/j.hcc.2024.100211

[99] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In 2018 IEEE 31st Computer Security Foundations Symposium (CSF) (Oxford, UK). IEEE, New York,

NY, USA, 268–282. doi:10.1109/CSF.2018.00027

[100] Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, Defu Lian, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2024. On the Vulnerability of Safety Alignment in Open-Access LLMs. In Findings of the Association for Computational Linguistics: ACL 2024 (Bangkok, Thailand), Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 9236–9260. doi:10.18653/v1/2024.findings-acl.549

[101] Ruisi Zhang, Seira Hidano, and Farinaz Koushanfar. 2022. Text Revealer: Private Text Reconstruction via Model Inversion Attacks against Transformers. arXiv:2209.10505 [cs.CL]

[102] Yiming Zhang, Nicholas Carlini, and Daphne Ippolito. 2024. Effective Prompt Extraction from Language Models. arXiv:2307.06865 [cs.CL] https://openreview.net/forum?id=0o95CVdNuz First Conference on Language Modeling (COLM 2024) (Philadelphia, PA, USA).

[103] Jian Zhao, Shenao Wang, Yanjie Zhao, Xinyi Hou, Kailong Wang, Peiming Gao, Yuanchao Zhang, Chen Wei, and Haoyu Wang. 2024. Models Are Codes: Towards Measuring Malicious Code Poisoning Attacks on Pre-trained Model Hubs. In Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering (Sacramento, CA, USA) (ASE '24). Association for Computing Machinery, New York, NY, USA, 2087–2098. doi:10.1145/3691620.3695271

[104] Weilin Zhong and Rezos. 2021. Code Injection. OWASP. Retrieved May 8, 2025 from https://owasp.org/www-community/attacks/Code_Injection

[105] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving Alignment and Robustness with Circuit Breakers. In Advances in Neural Information Processing Systems (Vancouver, BC, Canada), A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., Red Hook, NY, USA, 83345–83373. https://proceedings.neurips.cc/paper_files/paper/2024/file/97ca7168c2c333df5ea61ece3b3276e1-Paper-Conference.pdf

[106] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043 [cs.CL]

A. Terminology and Description 用語と説明

A.1 AI システムアーキテクチャのコンポーネントと関連用語

以下では図 2 に描かれた主要なコンポーネントと用語を説明し、想定される AI システムアーキテクチャ内での役割を明確にする。

- **AI モデル**：**AI システム**内にコンポーネントとして組み込まれる、人工知能技術に基づくモデル。この用語は、**事前学習済みモデル**と**ファインチューニング実施済みモデル**の両方を含む。
- **AI モデル開発者**：**学習環境**に関与する主体であり、**AI モデル**の開発のための **AI モデル情報**や学習データの提供、ファインチューニングなどの側面を担当する。
- **AI モデル情報**：**AI モデル**のアーキテクチャ、ハイパーパラメータ、学習アルゴリズム、および **AI モデル**の開発に必要なその他の技術仕様を含む、**AI モデル**の内部特性と構成を詳述する情報。
- **AI システム**：**AI モデル**をコンポーネントとして組み込んだ情報システム。
- **AI システム利用者**：**運用環境**でデプロイされた **AI モデル**とやり取りする主体（例：人間の利用者）。このやり取りには、システムへの入力（知識データを含む）の提供や、出力の受信が含まれうる。
- **アプリケーション**：**運用環境**内で、学習済み（ファインチューニング済み）の **AI モデル**がデータを処理して出力を生成するためにデプロイ、統合されるための枠組み。アプリケーションは、**内部知識**、**外部知識**、およびプラグインともやり取りする場合がある。
- **外部知識**：運用中にアクセスされる、**AI システム**の外部にある情報源。インターネットや**外部システム**など。これらの情報源はシステム内では維持されず、最新または補足的な情報を得るために必要に応じて照会される。
- **外部システム/センサー等**：**運用環境**で **AI システム**への入力データを提供する外部ソース、または **AI システム**の出力を受信する、もしくはその影響を受ける外部エンティティ。
- **ファインチューニング済み AI モデル**：**ファインチューニング用データセット**を使用し

たファインチューニングプロセスを経て、その能力を適応または特化させた AI モデル。このモデルはその後、通常は運用環境にデプロイされる。

- ファインチューニング用データセット：事前学習済み AI モデルのファインチューニングプロセスのために学習環境で特別に使用される、学習データの特定の集合。
- （運用環境への）入力（データ）：外部システム、センサー、または人間の利用者から提供され、運用環境内の AI モデルによって処理されるデータ。
- 内部知識：AI システムの利用者によって提供され、AI システム内に保存されるデータと文書の集合。アプリケーションは運用時に、プラグインや検索拡張生成（RAG; Retrieval Augmented generation）を介してこれらのデータにアクセスできる。
- 知識データ：内部文書や記録など、AI システムの利用者によって提供されるデータで、内部知識の一部として蓄積・保存されるものを指す。このデータは 1 回限りの入力として使用されるのではなく、将来の応答や決定をサポートするために保持される。
- 運用環境：学習済みの AI モデルがデプロイされ、入力を処理して出力を生成するためにアプリケーションフレームワークに統合される環境。
- （運用環境からの）出力（データ）：運用環境で AI モデルが入力を処理した後に生成される、潜在的な解釈情報を含む結果。
- 事前学習済み AI モデル：初期の学習を経た AI モデルであり、ファインチューニング用データセットを使用したさらなる専門的な学習（ファインチューニング）の基盤として機能する。
- 学習データ：学習環境内で AI モデルの学習またはファインチューニングに使用されるデータ。このデータは、実世界の運用を含む様々なソースから収集される場合がある。
- 学習データセット：学習環境内で AI モデルの初期学習またはその後のファインチューニングに使用される、学習データの特定の集合。
- 学習環境：AI モデルがデータセットを使用して学習およびファインチューニングを受ける環境。学習環境には、実世界の運用中に収集されたデータや、継続的または定期的なファインチューニングのシナリオを含む。

A.2 攻撃がおよぼす影響

以下のリストは、図 3 に描かれた AI システムへの攻撃に起因する各影響（1〜11）の詳細な説明であり、そのセキュリティ上の含意についてのさらなる明確化と包括的な理解を助ける。

(1) モデル漏洩：AI モデルのパラメータ、アーキテクチャ、ハイパーパラメータといった機

密性の高い内部詳細の不正な開示または抽出であり、攻撃者がモデルを複製または悪用することを可能にする。

(2) **学習データ漏洩**：機密性の高い学習データまたはそこから派生した情報の暴露または推測を指す。データのプライバシーと機密性を侵害する可能性がある。

(3) **アプリ情報漏洩**：アプリケーションの内部機能、設定、または関連するメタデータに関する機密情報の開示であり、さらなる攻撃を可能にする可能性がある。

(4) **入力情報漏洩**：利用者または外部システムによって入力として提供された機密データの不正な抽出または推測を指す。利用者のプライバシーまたは機密性を侵害しうる。

(5) **内部データ漏洩**：AI システム内に保存されている、または AI システムからアクセス可能な、組織の機密性の高い内部データの暴露または不正な抽出。

(6) **モデル誤動作**：AI モデルの挙動の変化または劣化であり、不正確な出力、誤解を招く出力、または劣化した出力をもたらし、信頼性と信用性を損なう。

(7) **解釈機能誤動作**：AI モデルの説明メカニズムの意図的な歪曲または破壊であり、モデルの意思決定プロセスの不正または誤解を招く解釈につながる。

(8) **セーフガード回避**：AI モデルの悪用または有害な挙動を防ぐために設計された保護措置の回避を指す。潜在的な不正行為につながる。

(9) **システム侵害**：AI システムの完全または部分的な乗っ取りを指す。攻撃者が不正なアクションを実行したり、権限を昇格させたり、後続のサイバー攻撃を容易にしたりといった可能性につながる。

(10) **内部データ毀損**：AI システムによって使用される、またはその内部に保存されている内部データの不正な変更、操作、または削除であり、データの完全性と可用性を侵害し、運用上の信頼性に影響を与える可能性がある。

(11) **計算資源浪費**：計算資源の過度の使用または過度の処理時間の意図的な誘発を指す。運用コストの増加、パフォーマンスの低下、またはサービス拒否状態を招く。