

機械学習システム セキュリティガイドライン

Version 1.02

2022年9月16日

機械学習システムセキュリティガイドライン策定委員会
機械学習システムセーフティ・セキュリティワーキンググループ[†]

日本ソフトウェア科学会 機械学習工学研究会



目次

Part-I. 「本編」

| | |
|---|------|
| I-1. ガイドライン概要 | I-1 |
| I-1.1. 本ガイドラインについて | I-1 |
| I-1.2. 目的・背景 | I-2 |
| I-1.3. スコープ | I-3 |
| I-1.3.1. 対象者とガイドラインの使われ方 | I-3 |
| I-1.3.2. 他の機械学習セキュリティ対策文献との位置づけ | I-4 |
| I-1.3.3. 一般的な情報セキュリティとの関連 | I-5 |
| I-1.4. 本ガイドラインで扱う機械学習システム | I-6 |
| I-2. 機械学習システム特有の攻撃 | I-7 |
| I-2.1. 機械学習システム特有の脅威 | I-7 |
| I-2.1.1. モデルやシステムの誤動作 | I-7 |
| I-2.1.2. モデルの窃取 | I-7 |
| I-2.1.3. 訓練データの窃取 | I-7 |
| I-2.2. 脅威を引き起こす攻撃 | I-8 |
| I-2.2.1. 回避攻撃 (evasion attack) | I-8 |
| I-2.2.2. ポイズニング攻撃 (poisoning attack) | I-8 |
| I-2.2.3. モデル抽出攻撃 (model extraction attack) | I-8 |
| I-2.2.4. モデルインバージョン攻撃 (model inversion attack) | I-8 |
| I-2.2.5. メンバシップ推測攻撃 (membership inference attack) | I-8 |
| I-3. 機械学習システムのセキュリティ | I-9 |
| I-3.1. 機械学習セキュリティの考え方 | I-9 |
| I-3.2. 進め方 | I-10 |
| I-3.3. 各工程の実施について | I-10 |
| I-4. 影響分析 | I-12 |
| I-4.1. 保護資産の特定 | I-12 |

| | |
|--|------|
| I-4.2. 関係する主体の整理..... | I-13 |
| I-4.3. 機械学習システム特有の脅威による影響分析..... | I-14 |
| I-5. システム仕様レベルでの脅威分析・対策..... | I-15 |
| I-5.1. システム仕様レベルでの脅威分析 | I-15 |
| I-5.1.1. 想定攻撃者の設定 | I-15 |
| I-5.1.2. 攻撃成立条件を満たすかの分析 | I-15 |
| I-5.2. 仕様レベルでの対策 | I-17 |
| I-6. 実際の機械学習システムに対する脅威分析・対策 | I-18 |
| I-6.1. 実モデルに対する脅威分析 | I-18 |
| I-6.2. 機械学習要素特有の対策..... | I-18 |
| I-7. 検知・対処 | I-20 |
| I-7.1. 機械学習システムセキュリティにおける検知・対処 | I-20 |
| I-7.2. 検知 | I-20 |
| I-7.3. 対処 | I-21 |
| I-7.3.1. 応急対策..... | I-22 |
| I-7.3.2. 調査 | I-23 |
| I-7.3.3. 恒久対策..... | I-24 |
| I-8. 参考文献..... | I-26 |
| Part-II. 「リスク分析編」 | |
| II-1. はじめに | II-1 |
| II-2. 本ガイドラインで扱う機械学習システムについて | II-2 |
| II-2.1. 機械学習システムの構成..... | II-2 |
| II-2.2. 機械学習システムの開発プロセス | II-3 |
| II-3. 機械学習システムセキュリティの概要..... | II-4 |
| II-3.1. 機械学習への攻撃..... | II-4 |

| | |
|--|-------|
| II-3.2. 攻撃による被害について | II-4 |
| II-4. 機械学習システムを守るには | II-6 |
| II-4.1. 機械学習システムを守る手段 | II-6 |
| II-4.2. 通常の IT セキュリティとの関係..... | II-6 |
| II-5. 機械学習システム開発プロセスにおけるリスク分析 | II-7 |
| II-5.1. 機械学習システム特有の攻撃に対するセキュリティを考慮した開発プロセス | II-7 |
| II-5.2. 機械学習システム向けの脅威分析技術について | II-9 |
| II-6. AI 開発者向けリスク分析..... | II-10 |
| II-6.1. AI 開発者向けリスク分析の概要..... | II-10 |
| II-6.2. 選択回答式 AI セキュリティリスク問診 (AI リスク問診) | II-10 |
| II-6.2.1. 機械学習セキュリティ専門家による事前準備手順 | II-11 |
| II-6.2.2. 分析者による分析手順 | II-14 |
| II-7. AI リスク問診の実現例 | II-17 |
| II-7.1. 注意事項 | II-17 |
| II-7.2. アタックツリーと攻撃実施可能条件 | II-17 |
| II-7.2.1. 回避攻撃（敵対的サンプル）のアタックツリーと攻撃実施好条件..... | II-19 |
| II-7.2.2. ポイズニング攻撃のアタックツリーと攻撃実施可能条件 | II-23 |
| II-7.2.3. モデル抽出攻撃のアタックツリーと攻撃実施可能条件 | II-26 |
| II-7.2.4. モデルインバージョン攻撃のアタックツリーと攻撃実施可能条件..... | II-30 |
| II-7.2.5. メンバシップ推測攻撃のアタックツリーと攻撃実施可能条件 | II-32 |
| II-7.3. 質問群..... | II-37 |
| II-7.4. 攻撃実施可能条件の満足状況の判定用テーブル | II-41 |
| II-7.5. AI リスク問診ツール | II-44 |
| II-8. AI リスク問診の試行例 | II-45 |
| II-8.1. 事例試行概要..... | II-45 |
| II-8.1.1. 融資審査 AI | II-45 |
| II-8.1.2. プラント制御 AI..... | II-62 |

| | |
|----------------------------|-------|
| II-8.1.3. 性別・年齢推定 AI | II-67 |
| II-9.まとめ | II-91 |
| II-10.参考文献..... | II-92 |
| 「付録：攻撃検知技術の概要」 | |
| A-1.はじめに..... | A-1 |
| A-2.攻撃戦略毎の検知手法について | A-1 |
| A-2.1.前兆検知 | A-1 |
| A-2.1.1.回避攻撃の検知..... | A-1 |
| A-2.1.2.ポイズニング攻撃の検知 | A-2 |
| A-2.2.兆候検知 | A-3 |
| A-2.2.1.回避攻撃の検知..... | A-3 |
| A-2.2.2.ポイズニング攻撃の検知 | A-3 |
| A-2.2.3.モデル抽出攻撃の検知 | A-3 |
| A-3.検知に使用するデータについて | A-3 |
| A-4.まとめ | A-5 |
| A-5.参考文献..... | A-6 |

機械学習システムセキュリティガイドライン策定委員会メンバーリスト131

機械学習システムセキュリティガイドライン

機械学習システム
セキュリティガイドライン
Part I. 「本編」

Version 1.02

2022 年 9 月 16 日

機械学習システムセキュリティガイドライン策定委員会
機械学習システムセーフティ・セキュリティワーキンググループ[†]

日本ソフトウェア科学会 機械学習工学研究会



I-1. ガイドライン概要

I-1.1. 本ガイドラインについて

本書は、機械学習システムの開発者・サービス提供者向けに、機械学習システム特有の攻撃に対するセキュリティ対策手順を整理したものである。セキュリティ対策の実施において、すべきことの把握や、実施時の開発者・サービス提供者と機械学習セキュリティ専門家との意志疎通を助けることを目的とした資料である。本書は「法的拘束力のないガイドライン」であり強制力はない。

本書は、「本編」「リスク分析編」と「攻撃検知技術の概要」で構成される（図 I-1）。

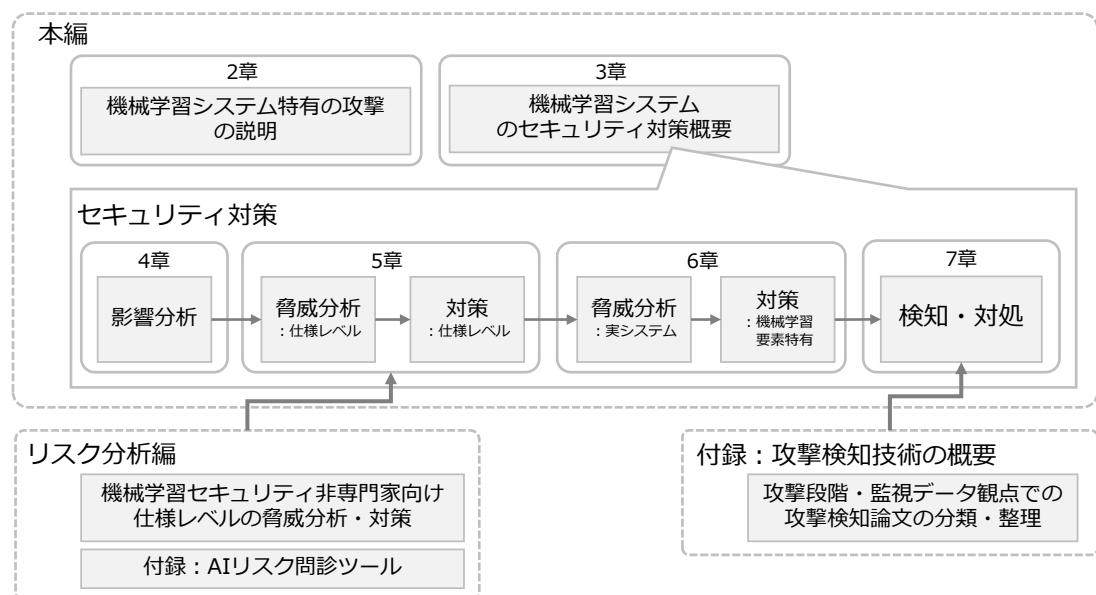


図 I-1. ガイドライン概要

本編では、機械学習システム特有の攻撃に対するセキュリティ対策の手順を整理している。I-2 章で機械学習システム特有の攻撃について説明し、I-3 章で機械学習システム特有の攻撃へのセキュリティ対策手順の概要や実施者について説明する。I-4～I-7 章では I-3 章で紹介された各手順について実施する内容を説明する。

リスク分析編では、本編のうち、I-5 章で説明する「システム仕様レベルでの脅威分析・対策」を、機械学習セキュリティの専門知識がないシステム開発者自身で分析する手法について説明する。

付録の「攻撃検知技術の概要」では、I-7 章で説明する検知・対処について、機械学習システム特有の攻撃に対する検知技術論文を、「対象とする攻撃」と「その攻撃段階」や「監視するデータ」で分類・整理したものであり、実際に検知システムを構築する際の参考とな

る。

I-1.2. 目的・背景

近年、機械学習の発展とその普及に伴い、機械学習を活用し、画像認識や自然言語処理などこれまで実現できなかった高度な機能をもつシステム（以下、AI システム）が広く開発されるようになってきている。そのような AI システムが自動運転や金融取引の自動化などに使われ、社会インフラとして欠かせないものになると、セキュリティ上の被害が大きくなる可能性が高くなり、その考慮が必要になる。

機械学習を使わない従来のシステムであってもセキュリティの分析や対策は行われてきたが、機械学習には、訓練済みモデルを故意に誤動作させる敵対的サンプルなど、機械学習システム特有の脆弱性が発見されており、機械学習システム特有のセキュリティ分析や対策が必要となる。

2014 年に敵対的サンプルが発見されて以来、機械学習システム特有の攻撃やその防御方法については、盛んに研究がなされている。しかしながら、各研究で提案されている攻撃や防御についての前提条件や評価指標はまちまちで、研究動向を整理した論文はあるものの、実際の開発において、どのような攻撃の可能性があり、対策をどの程度考慮する必要があるかは、必ずしも明らかになっていない。さらに、機械学習システム特有のセキュリティの分析や対策の実施は、機械学習とセキュリティの両方の知識を必要とし、この両方の知識をもつ機械学習セキュリティの専門家が少ない開発現場では、実施困難な活動となる。

そこで、専門知識を持たない AI 開発者が、構築する AI に機械学習システム特有の攻撃が起こりうるかを判断する基準（ガイドライン）が必要になる。本ガイドラインでは、典型的な攻撃手法を整理し、その攻撃を実行できる条件が開発中の AI において満たされるかどうか AI 開発者が判断できるようにしている。この基準に照らし合わせて、攻撃の可能性を分析することにより、適宜、機械学習セキュリティの専門家と連携すべきかどうかを判断できるようになる。

このガイドラインを整理するにあたり、機械学習工学研究会 機械学習システム セーフティ・セキュリティワーキンググループ内に、機械学習セキュリティガイドライン策定委員会を 2021 年 7 月に設置した。本策定委員会は、本研究会でメンバーを公募し、それに応じた産学の委員で構成され、AI 開発者にとって有益な情報をガイドラインとして提供することを目指している。本ガイドラインは、本策定委員会での議論、検討の結果をまとめている。

AI の開発・利活用に関して、政府や国際機関がガイドラインを公表している。2019 年には OECD から複数国で合意された AI 原則[I-1]が公表され、包摂的な成長、持続可能な開発及び幸福、人間中心の価値観及び公平性、透明性及び説明可能性、堅牢性、セキュリティ及び安全性、アカウンタビリティが挙げられている。日本では「人間中心の AI 社会原則」[I-2]において、人間中心の原則、教育・リテラシーの原則、プライバシー確保の原則、セキュリティ確保の原則、公正競争確保の原則、公平性、説明責任及び透明性の原則、イノベー

ションの原則を掲げている。

AI 原則を構成する諸要素のまとめ方はそれぞれ異なるが、プライバシー、アカウンタビリティ、安全性とセキュリティ、透明性と説明可能性、公正性と非差別性、人間による技術管理、専門家の責任、人間的な価値の促進からなる 8 つのテーマに区分されるという。

上記の原則を実践するための方策である AI ガバナンスの構造を経済産業省の「我が国の AI ガバナンスの在り方」[I-3]では図 I- 2 のように整理している。

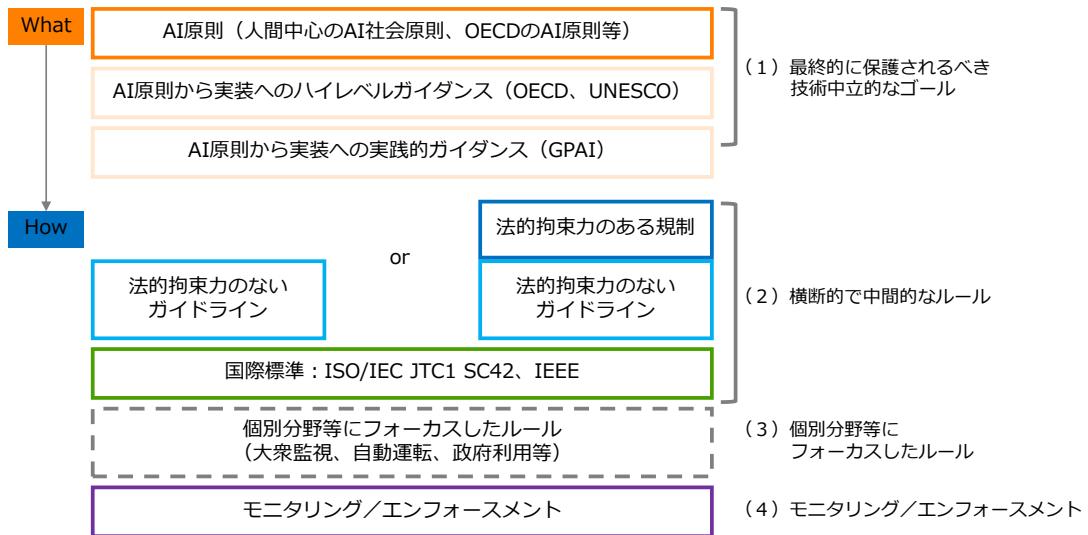


図 I- 2 AI ガバナンスの構造

本ガイドラインは図 I- 2 の「法的拘束力のないガイドライン」にあたり、AI 原則でとりあげられる「セキュリティ」を尊重するための取り組みである。

I-1.3. スコープ

本節では本ガイドラインが想定する読者とガイドラインの使われ方、そして他の関連文献との位置づけを説明する。

I-1.3.1. 対象者とガイドラインの使われ方

本ガイドラインが想定する読者は、機械学習セキュリティの専門知識をもたない、機械学習システムの開発者や機械学習システムを利用したサービスの提供者である。機械学習システムのセキュリティ対策において、機械学習セキュリティの専門家と非専門家で行うことと切り分けて説明する。

総務省の AI 利活用ガイドライン[I-4]において、AI の利活用の流れが図 I- 3 のように整理されている。

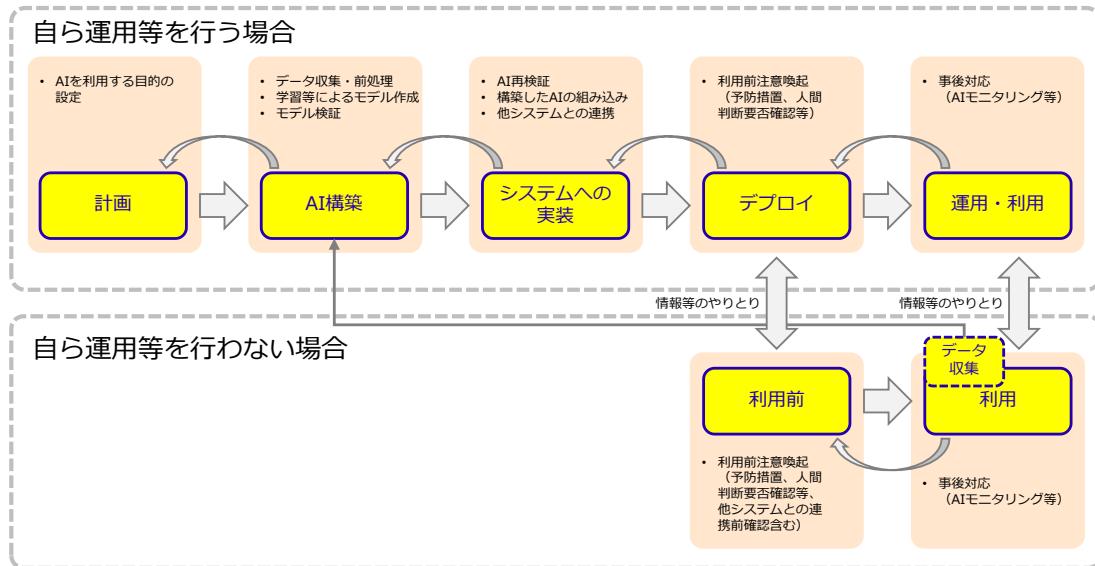


図 I-3 AI の利活用の流れ

本書は、「計画」や「AI 構築」、「運用・利用」で利用されることを想定している。

I-1.3.2. 他の機械学習セキュリティ対策文献との位置づけ

機械学習システム特有の攻撃へのセキュリティ対策に関する文献は、国内外の政府機関、企業から様々公開されている。

MITRE が公開している ATLAS[I-5]では、機械学習システム特有の攻撃の戦術・手法の整理を行っており、偵察や初期アクセスから攻撃の実行、情報窃取等の攻撃段階毎に使用される手法をまとめている。また、実際の攻撃事例をケーススタディとしてまとめている。

機械学習システムのリスク分析については、ENISA の Artificial Intelligence Cybersecurity Challenges [I-6]で機械学習システムが関連しうる資産や発生する被害について触れられている。また、Microsoft の Threat Modeling AI/ML Systems and Dependencies[I-6]や ICO の AI and data protection risk mitigation and management toolkit[I-8]では AI システム開発時に気にするべき点について触れられており、開発しているシステムと脅威との結びつけを助ける資料となっている。

機械学習システム特有の攻撃とその防御策については、NIST[I-9]、産総研[I-10]、ENISA[I-6]、ICO[I-8]等の機関や Microsoft[I-6]、三井物産セキュアディレクション株式会社[I-11]等の企業が公開を行っている。

一方で、既存の機械学習システム特有の攻撃へのセキュリティ対策文献には、一連のセキュリティ対策を体系的に整理したものはなく、実際にシステム開発者やサービス提供者がセキュリティ対策を実施するのが困難であった。そこで本ガイドラインでは、セキュリティ対策実施における、すべきことの把握や、実施時の開発者・サービス提供者と機械学習セキュリティ専門家との意志疎通を助けるために、機械学習システムの開発者やサービス提供

者向けに、セキュリティ対策手順の整理を行った。

また、セキュリティ対策実施のハードルを下げるため、セキュリティ対策の手順における「システム仕様レベルでの脅威分析・対策」を、機械学習セキュリティの専門知識がなくとも実施できる手法を提案する。

I-1.3.3. 一般的な情報セキュリティとの関連

本ガイドラインで扱う機械学習システム特有の攻撃に対するセキュリティと一般的な情報セキュリティとの関連について説明する。本ガイドラインでは、機械学習システム独自の特性を利用し、正規の権限で可能なアクセスを用いた攻撃を「機械学習システム特有の攻撃」と定義する。例えば、機械学習システムへ入力を繰り返し、その出力を観察することで、対象の機械学習システムのモデルを複製する攻撃等が含まれる。セキュリティ対策においては、一般的な情報セキュリティ・機械学習システム特有の攻撃に対するセキュリティ、両方が行われる必要がある。対策においては、攻撃による被害や攻撃の発生可能性を参考に、一般的な情報セキュリティ・機械学習システム特有の攻撃に対するセキュリティ両方の優先度をあわせて検討することが望ましい。一般的な情報セキュリティのリスク分析や対策については ISO 27000 シリーズ、ISO/IEC 15408、NIST SP 800 シリーズ、NIST Cyber Security Framework、情報処理推進機構のガイドライン等のフレームワークが提供されている。

I-1.4. 本ガイドラインで扱う機械学習システム

本ガイドラインで対象とする機械学習システムは、機械学習(Machine Learning)を用いたシステムである。機械学習システムの機械学習処理部は訓練パイプラインと推論パイプラインから構成されるのが一般的であり、図 I-4・図 I-5 のように表すことができる。システムによっては訓練パイプラインを外部で行い、推論パイプラインのみ行う場合もある。機械学習システムの運用に先立って、訓練パイプラインにて訓練関連データを用いて訓練処理を行い、訓練済みモデルを生成する。そして、推論パイプラインにて推論対象データと訓練済みモデルを用いて推論処理を行い、推論結果を得る。

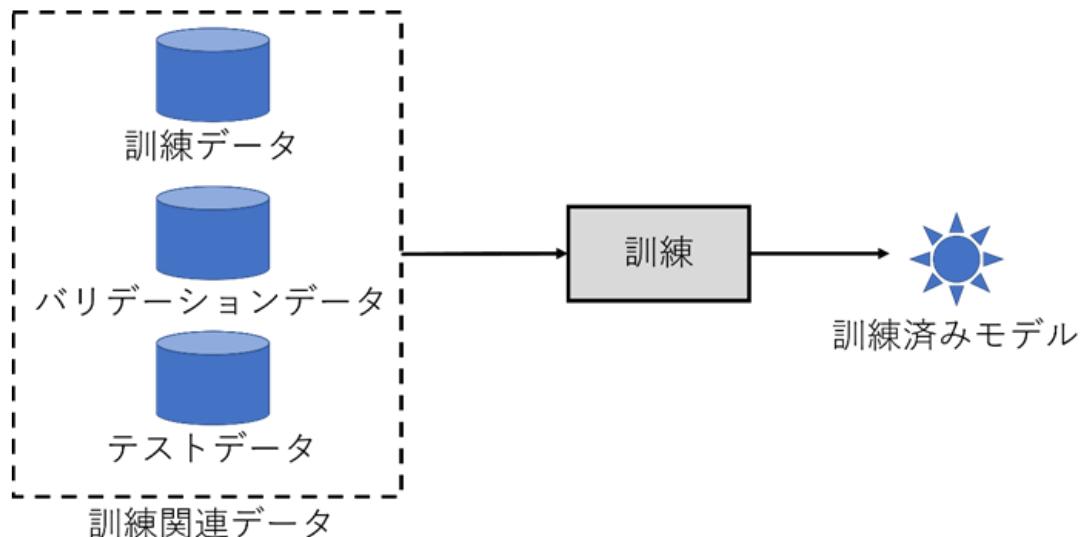


図 I-4 機械学習処理部の訓練パイプライン

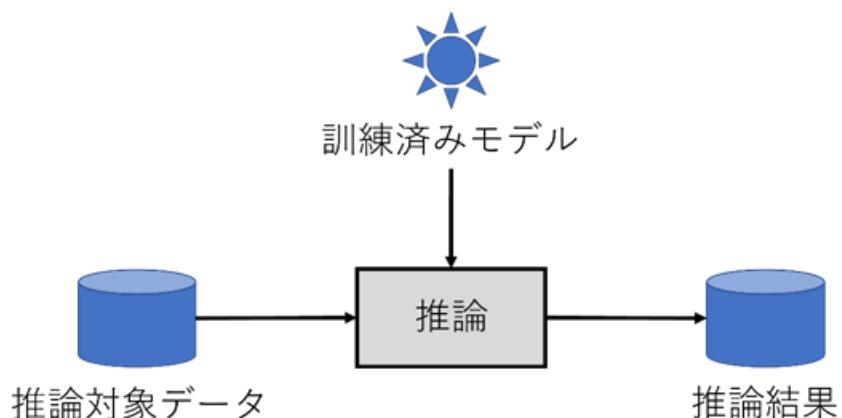


図 I-5 機械学習処理部の推論パイプライン

I-2. 機械学習システム特有の攻撃

本章では機械学習システム特有の攻撃による脅威、脅威を引き起こす具体的な攻撃について説明し、最後にセキュリティ対策の基本的な考え方を示す。

機械学習システム特有の攻撃については、NIST の NIST IR 8269 Draft[I-9]や産総研の機械学習品質マネジメントガイドライン[I-10]、Microsoft の公開している Threat Modeling AI/ML Systems and Dependencies[I-6]においても整理されており、参考となる。

I-2.1. 機械学習システム特有の脅威

機械学習システムに特有の脅威は、大きく以下の3つに分類できる。これらは、機械学習システムに対する正規の権限でのアクセスによって引き起こされる可能性がある。

- モデルやシステムの誤動作
- モデルの窃取
- 訓練データの窃取

I-2.1.1. モデルやシステムの誤動作

本脅威は、機械学習システムのモデルを誤動作させられ、システムに本来期待される動作が阻害される脅威である。モデルやシステムの誤動作は、自動運転システムの事故誘発（標識認識システムを誤認識させる等）、マルウェア検知の回避等の被害につながりうる。

モデルやシステムの誤動作を引き起こす攻撃は、推論時のモデル・システムへの入力によって行う攻撃（I-2.2.1 項）や訓練データ・訓練モデルを汚染する攻撃（I-2.2.2 項）がある。また、判定結果を意図的に変えるものだけでなく、モデルやシステムの説明性機能だけを誤動作させるものもあり、システムの透明性に影響を与える可能性がある。

I-2.1.2. モデルの窃取

本脅威は、機械学習システムのモデルのコピー、または、近い性能のモデルを攻撃者に作成されるという脅威である。モデルの窃取は、サービスを複製される等の直接的な知財への被害以外に、窃取したモデルを利用した他の攻撃への被害にもつながりうる。

モデルの窃取を行う攻撃は、モデルやシステムへの入力によって行う攻撃（I-2.2.3 項）がある。

I-2.1.3. 訓練データの窃取

本脅威は、機械学習システムのモデルの訓練に使われたデータそのもの、または、訓練に使われたデータの情報の一部を攻撃者に推測されるという脅威である。被害として個人情報の漏洩等のプライバシー保護への被害につながりうる。

訓練データの窃取を行う攻撃は、モデルやシステムへの入力によって行う攻撃（I-2.2.4 項、I-2.2.5 項）がある。

I-2.2. 脅威を引き起こす攻撃

I-2.1 節で挙げた脅威を引き起こす機械学習システム特有の攻撃として代表的な5つの攻撃を挙げる。

I-2.2.1. 回避攻撃 (evasion attack)

本攻撃はモデルやシステムの誤動作 (I-2.1.1 項) を引き起こす攻撃である。

機械学習システムへの入力に悪意のある変更を加えることで、システムが意図していない動作をさせる。敵対的サンプル (adversarial example) と呼ばれる、入力データに人間にはわからないくらいわずかなノイズを加えることでモデルの誤判断を誘発する攻撃が有名である。

I-2.2.2. ポイズニング攻撃 (poisoning attack)

本攻撃はモデルやシステムの誤動作 (I-2.1.1 項) を引き起こす攻撃である。

攻撃者が、細工したデータ・モデルを、機械学習システムのモデルの訓練に利用されるデータやモデルに紛れ込ませることで誤動作させる。特定のラベルを別のラベルに誤判定させる攻撃だけでなく、トリガーと呼ばれる特定のパターンが含まれた入力を特定のラベルに誤判断させるバックドア攻撃がある。

I-2.2.3. モデル抽出攻撃 (model extraction attack)

本攻撃はモデルの窃取 (I-2.1.2 項) を引き起こす攻撃である。

機械学習システムへの入力に対する出力を分析することで、対象システムのモデルと同等の性能をもつモデルを作成する攻撃である。

I-2.2.4. モデルインバージョン攻撃 (model inversion attack)

本攻撃は訓練データの窃取 (I-2.1.3 項) を引き起こす攻撃である。

機械学習システムへの入力に対する出力を分析することで、訓練データに含まれる情報を復元する攻撃である。

I-2.2.5. メンバシップ推測攻撃 (membership inference attack)

本攻撃は訓練データの窃取 (I-2.1.3 項) を引き起こす攻撃である。

機械学習システムへの入力に対する出力を分析することで、ある対象のデータがモデルの訓練データに含まれているかを特定する攻撃である。モデルインバージョン攻撃と異なり、訓練データ自体を復元するものではない。

I-3. 機械学習システムのセキュリティ

本章では機械学習システム特有の攻撃へのセキュリティ対策の手順について説明する。機械学習システムにおいては、これまでに對応してきた一般的な情報セキュリティに加え、機械学習システム特有の攻撃に対するセキュリティ対策を行う必要がある。対策においては、攻撃による被害や攻撃の発生可能性を参考に、一般的な情報セキュリティ・機械学習システム特有の攻撃に対するセキュリティ両方の優先度をあわせて検討することが望ましい。

なお、セキュリティ対策においては、通常、すべての攻撃を網羅的に把握し対策することができない点には留意が必要である。

I-3.1. 機械学習セキュリティの考え方

基本的には一般的な情報セキュリティ対策と同様に、攻撃の可能性を洗い出してその対策を行う、「リスク分析」⇒「対策」のプロセスで考える。リスク分析は、影響分析と脅威分析からなる。まず、システムが関連する保護資産を特定し、保護資産に影響を与える脅威（I-2.1 節）と結びつけ、脅威が発生した際の影響を分析する。次に、脅威を引き起こす攻撃（I-2.2 節）が可能となる脆弱性がシステムにあるかの脅威分析を行う。その後、影響分析と脅威分析の結果をふまえてリスクを算出する。

機械学習における保護資産・脅威・攻撃の関連と、影響分析・脅威分析、対策検討のイメージを図 I- 6 に示す。

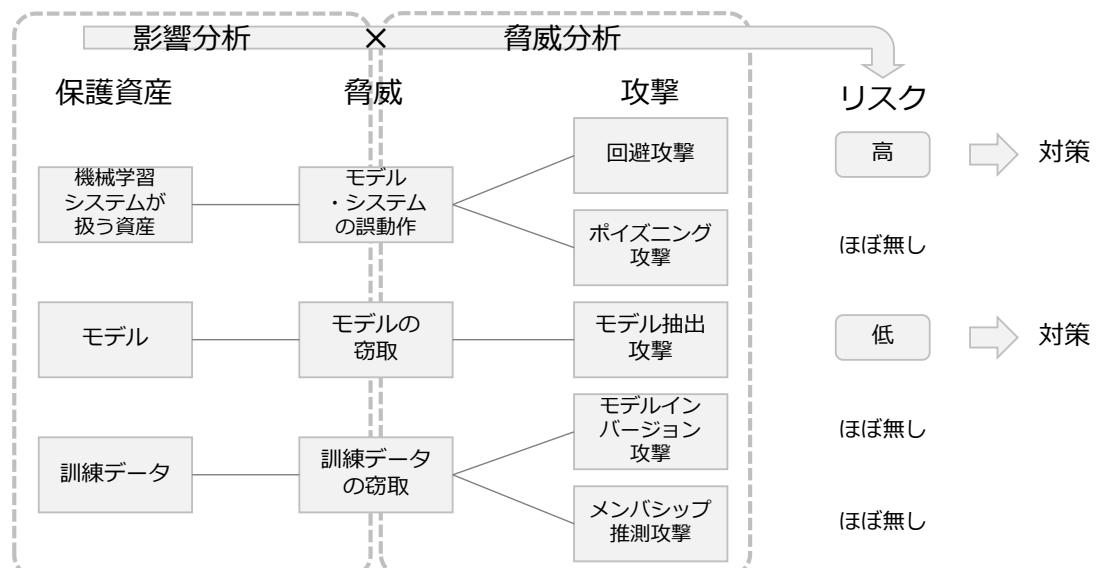


図 I- 6 機械学習セキュリティの全体イメージ

対策は、運用・利用前に攻撃を困難にしたり、攻撃の効果を抑制したりする「緩和策」と、運用中の攻撃発生を見つけ、対応する「検知・対処」の大きく2つにわけられる。基本的に

は、可能な限り運用前の緩和策によって対策を実施し、対策できなかった脅威やより重点的に対策する脅威、事前に洗い出せなかつた未知の脅威への対策として、「検知・対処」を実施する。

I-3.2. 進め方

前節各工程の進め方の例を図 I-7 に示す。

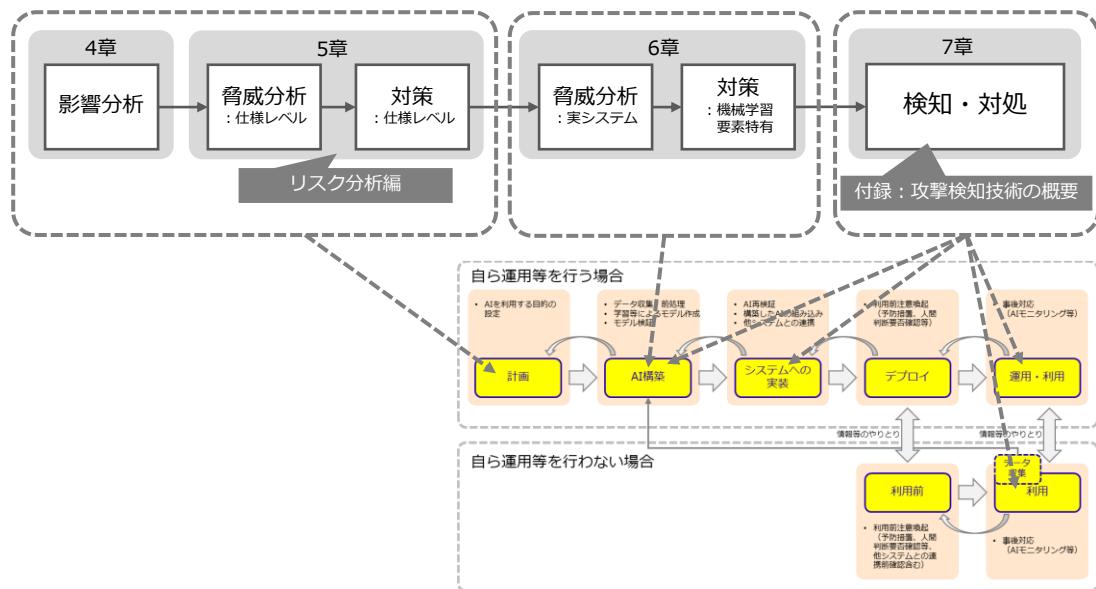


図 I-7 機械学習セキュリティの工程

まず、「影響分析」を行い、次に「システム仕様レベルでの脅威分析・対策」を行う。この段階で脅威への対処が十分でなければ、「実際の機械学習システムに対する脅威分析・対策」を行う。その後、対策結果をうけて、未対策脅威への対策や安全策として、「検知・対処」を設計、実施する。実際の検知・対処については、運用・利用時が基本であるが、汚染されたデータ・モデルの混入対策等、AI 構築やシステムへの実装段階でも行われる場合がある。その場合は「計画」の段階で「AI 構築」時の「検知・対処」体制を整えておく必要がある。

本書では、上記工程のうち、「システム仕様レベルでの脅威分析・対策」について、機械学習セキュリティの専門知識がないシステム開発者自身で分析する手法を「リスク分析編」で説明する。また、検知・対処について、機械学習システム特有の攻撃に対する検知技術論文を、「対象とする攻撃」と「その攻撃段階」で分類・整理したものを付録の「攻撃検知技術の概要」で紹介する。

I-3.3. 各工程の実施について

前節の各工程の実施においては、実際に分析や対策を行う実施者と、各工程の間で対策や検知・対処の実施を決める責任者が必要であり、両者の協力が不可欠である。

実施者は基本的にシステム開発者となる。ただし、検知・対処については、ログ設計等、事前の準備は計画～システムへの実装で行われるが、実際の検知・対処は基本的に運用・利用フェーズで実施される。そのため、前者はシステム開発者、後者はサービス提供者によって実施される可能性が高く、両者の間で意思疎通が必要である。各工程の実施には機械学習セキュリティの専門知識が不可欠であり、専門家との協力が必要となる。

責任者はシステム開発の依頼元やサービス提供者といった、リスクが顕在化した際にその責任を取る者（リスクオーナー）となる。実施者より提示される影響分析や脅威分析の結果をみて、対策すべき脅威やその優先度、制約条件を決定する。

I-4. 影響分析

本章では、機械学習システム特有の攻撃に対する影響分析について説明する。手順としては、機械学習システムの保護対象となる資産を特定し、特定された資産と機械学習システム特有の攻撃によって引き起こされる脅威とを結びつけ、想定される被害を算出する。

I-4.1. 保護資産の特定

機械学習システムが関連する資産を特定する。機械学習システムが関連する資産は、モデル・訓練データといった「機械学習を構成する資産」と、モデルの出力結果によって影響をうける「機械学習システムが扱う資産」の大きく2つがある。

各システムにおける「機械学習システムが扱う資産」の一例を以下に挙げる。

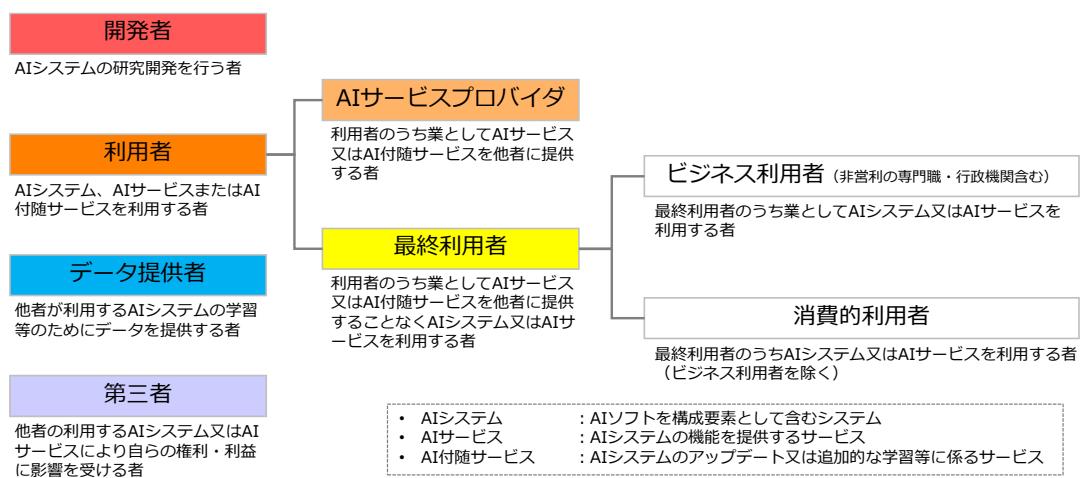
- 標識認識システム ⇒自動運転
- レントゲン画像診断システム ⇒医療判断
- 顔認証ゲート ⇒ゲート設置場所のセキュリティ
- SNS 画像フィルタ ⇒SNS ポリシー
- マルウェア検知 ⇒設置組織・端末の情報セキュリティ

機械学習システムが関連する資産(アセット)については、ENISA の Artificial Intelligence Cybersecurity Challenges [I-6]でも触れられている。主に「機械学習を構成する資産」について、データ・モデル・ステークホルダー等6つのカテゴリで分類・列挙されており、保護資産を特定する際の参考となる。

I-4.2. 関係する主体の整理

保護資産に結びつく関係者を整理し、脅威によって影響をうける主体を明確化する。例えば、訓練データであればデータの提供者が関係者にあたる。「機械学習システムが扱う資産」については、システム利用者以外に影響が及ぶ可能性があることも留意する。例えば、レントゲン画像診断システムの場合、システムの直接の利用者は医師であるが、システムの判定結果によって影響を受けるのは、医師だけでなく診断される患者も含まれる。

AI利活用ガイドライン[I-4]ではAIの利活用において関与が想定される主体を図I-8のように整理しており、参考になる。



図I-8 AIの利活用において関与が想定される主体

I-4.3. 機械学習システム特有の脅威による影響分析

前節までに特定した保護資産・関係主体と、I-2.1 節の脅威とを結びつけ、発生しうる被害を分析する。結びつけを図 I-9 に示す。

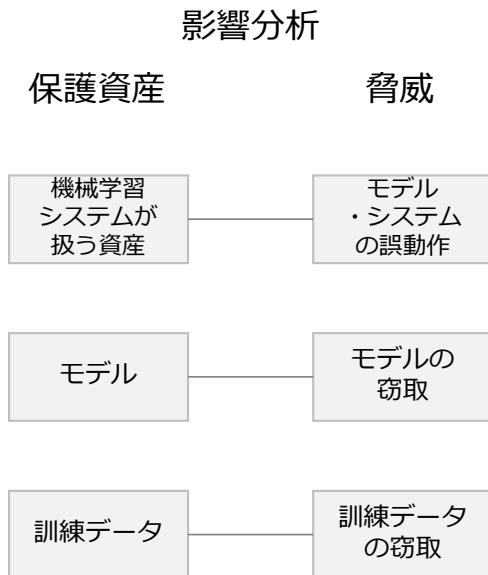


図 I-9 保護資産と脅威の結びつけ

「機械学習システムが扱う資産」と「モデル・システムの誤動作」の結びつけについては、具体的な脅威シナリオの想定が必要となる。例を以下に挙げる。

- 標識認識システムを誤認識させることで自動運転車の事故を引き起こす
- レントゲン画像判断システムを誤判断させることで、医療ミスを引き起こす
- 顔認証ゲートを誤判断させることで、登録者以外のゲート通過を許してしまう
- SNS の画像フィルタを回避することで、規約で禁止された画像をアップロード・公開する
- マルウェア検知を回避し、対象者にマルウェアをダウンロードさせる

「機械学習システムが扱う資産」の影響分析として、産総研の機械学習品質マネジメントガイドライン[I-10]では人に対する傷害などの人的リスクとその他の経済的リスクに細分し、7 レベルの「AI 安全性レベル」を定義しており、そちらも参考となる。

I-5. システム仕様レベルでの脅威分析・対策

本章では機械学習システムの仕様レベルでの脅威分析・対策について説明する。総務省のAI利活用ガイドライン[I-4]のAI利活用の流れにおける、「計画」フェーズ（図I-3）で行うことを見定している。

I-4章の影響分析の結果をうけて、想定される被害を看過できない攻撃に対して脅威分析を行う。仕様レベルで分析・対策することで、後の「AI構築」・「システムへの実装」時の実際のシステムに対する脅威分析・対策において、対象とする攻撃の絞り込みや、分析・対策自体の省略が可能となる。

なお、本章の具体的な実施を、機械学習セキュリティの専門知識を持たないシステム開発者やサービス提供者でも可能にした手法を「リスク分析編」にまとめている。

I-5.1. システム仕様レベルでの脅威分析

機械学習システム特有の脅威を引き起こす攻撃（I-2.2節）について、対象の機械学習システムにおいて、システム仕様上、攻撃実行が可能かを分析する。手順としては、まず対象の機械学習システムに関する主体（I-4.2節）の中から想定攻撃者を設定し（I-5.1.1項）、次に、想定攻撃者の能力が対象とする攻撃の成立条件を満たすかを分析（I-5.1.2項）する。各工程について説明する。

I-5.1.1. 想定攻撃者の設定

まず、対象の機械学習システムに関する主体（I-4.2節）の中から想定攻撃者を設定する。想定攻撃者とは、対象の機械学習システムに関する主体の中で、攻撃者となりうる関係者である。「システム運用者は訓練処理を実施できるが、システム利用者は推論処理しか実施できない」等、関係者毎に対象の機械学習システムにおける権限が異なるため、想定攻撃者を複数設定し、想定攻撃者毎にI-5.1.2項の攻撃成立条件を満たすかの分析を行う。

想定攻撃者の設定においては、「システムの管理・運用者（開発者・サービス・プロバイダ）」「訓練データの提供者」「システムの利用者」「システムを直接利用しない第三者」の4つの観点が考えられる。例えば、レントゲン画像の病理判定システムの場合、「システムの管理者」「訓練用のレントゲン画像データ提供者」「医者（システムの利用者）」「患者（システムを直接利用しない第三者）」が想定攻撃者の候補となる。AI利活用ガイドライン[I-4]ではAIの利活用において関与が想定される主体を図I-8のように整理しており、想定攻撃者を網羅的に洗い出す際の参考となる。

I-5.1.2. 攻撃成立条件を満たすかの分析

次に、想定攻撃者の能力が攻撃成立条件を満たすか分析する。

I-2.2節で挙げた5つの攻撃について、基本的には、機械学習システムへの推論処理と推論結果取得の回数が攻撃成立条件となり、ポイズニング攻撃は加えて訓練処理や訓練データ

タ・モデルへの介入度合いが条件に追加される。各攻撃の成立条件を緩和する要素として、取得できる推論結果の内容、訓練データ・類似データの入手、システムに関する公開情報等が関連する。

具体的にどれだけ訓練・推論が実施できれば攻撃が成立するかは最新の研究動向を調査し算出する必要がある。仕様レベルにおいては、対象の機械学習システムと同じ機械学習アルゴリズム、類似のデータを扱う研究等を参考に設定する。各観点における想定攻撃者の能力の判断要素例を表 I- 1 に挙げる。

表 I- 1. 想定攻撃者の能力の判断要素例

| 判断要素 | 例 |
|------------------|---|
| 訓練への介入度合い | <ul style="list-style-type: none"> ➢ 訓練データへ任意のデータをどれくらい混入することができるか ➢ 訓練に利用されるモデルへ任意のモデルをどれくらい混ぜることができるか ➢ 訓練済みモデルを直接置き換えられる場合（例：想定攻撃者がモデル訓練を委託される立場）、推論に関する条件は関係なくポイズニング攻撃は可能 |
| 推論処理・推論結果取得の回数 | <ul style="list-style-type: none"> ➢ 訓練済みモデルを入手できるなら（例：車載システム）自由に推論処理を実行できる ➢ 取得できる推論結果の内容 <ul style="list-style-type: none"> ◆ ラベルだけでなく確信度も得られるか ◆ 推論時のモデルの内部出力を得られるか ◆ 直接推論結果が得られなくとも、システムの挙動で推論結果を推測可能な場合がある（例：標識認識システム） |
| 関連データの入手 | <ul style="list-style-type: none"> ➢ 訓練データの一部やその類似データ、訓練データの統計情報等があるとモデル抽出攻撃・モデルインバージョン攻撃・メンバシップ推測攻撃が容易になる |
| 機械学習システムに関する公開情報 | <ul style="list-style-type: none"> ➢ 使用する機械学習アルゴリズム ➢ 入力データの仕様 ➢ モデル入力の前処理 ➢ 訓練時のパラメータ |

分析の際には、想定攻撃者の能力が攻撃成立条件を満たすかだけでなく、実際に攻撃成立可能なくらい推論処理を実行するにはどれだけの時間がかかるか等、条件を満たすのが現

実的かも加味し、攻撃可能性を算出する。

I-5.2. 仕様レベルでの対策

I-4 章の影響分析の結果と I-5.1 節の脅威分析の結果を組み合わせ、対策が必要な攻撃について、仕様レベルの対策を実施する。基本的には、想定被害が大きく、実行が容易である攻撃への対策を優先する。

仕様レベルの対策としては主に、システム全体での緩和策と開発プロセスにおける緩和策があり、I-5.1 節で分析した攻撃条件を想定攻撃者が満たせなくなるように実施する。また、そもそもシステムの利用方法を変更することで被害自体を緩和する対策も考えられる。表 I- 2 に各緩和策の例を示す。

表 I- 2 仕様レベルの緩和策

| 緩和策の分類 | 例 |
|---------------|--|
| システム全体での緩和策 | <ul style="list-style-type: none"> ➤ 想定攻撃者がシステム・モデルへ入力できる機会を減らす <ul style="list-style-type: none"> ◆ 入力回数に制限をつける ➤ 必要以上の情報を出力しない <ul style="list-style-type: none"> ◆ 判定ラベルの上位 1 件のみ、等出力を制限する ◆ 判定ラベルのみ出力し、確信度は出力しない ➤ システム・モデルに関する公開情報を減らす ➤ 不特定多数が訓練データを入力できないようにする |
| 開発プロセスにおける緩和策 | <ul style="list-style-type: none"> ➤ 訓練を利用するデータ・モデルは信頼できる提供者のものを利用する |
| 被害自体の緩和策 | <ul style="list-style-type: none"> ➤ リスクを許容するシステム設計に変更する <ul style="list-style-type: none"> ◆ 例：モデルの出力から次の処理を自動実行するシステムにおいて、出力と次の処理の間で人間の判断を追加する |

仕様レベルで対策を行えない場合、次章の実際の機械学習システムに対する脅威分析・対策を行う。

I-6. 実際の機械学習システムに対する脅威分析・対策

本章では実際に機械学習システムに対して行う脅威分析・対策について説明する。総務省のAI利活用ガイドライン[I-4]のAI利活用の流れにおける、「AI構築」フェーズで行うことと想定している。

I-5章のシステム仕様レベルでの脅威分析・対策において、対策ができなかった攻撃(I-2.2節)について、攻撃可能性を実際に評価し、機械学習要素に特有の対策を実施する。

I-6.1. 実モデルに対する脅威分析

機械学習システム特有の脅威を引き起こす攻撃(I-2.2節)について、対象の機械学習システムにおいて実際に攻撃実行が可能かを分析する。

まず、対象の攻撃(I-2.2節)について、対象の機械学習システムの機械学習アルゴリズムや扱うデータの種類を元に、発表されている論文等から適用可能な具体的な攻撃手法を抽出する。次に、抽出した攻撃手法を実装し、攻撃がどの程度成功するか評価する。攻撃手法の実施を補助するツールとしては、Adversarial Robustness Toolbox(ART)[I-12]、CleverHans[I-13]、Counterfit[I-14]などが公開されている。

I-6.2. 機械学習要素特有の対策

前節の結果より、対策が必要な攻撃について、機械学習要素特有の緩和策を実施する。緩和策の例を攻撃毎に表I-3に示す。緩和策実施後は、再度I-6.1節の脅威分析を行い、効果を評価・確認する。

表 I- 3 機械学習要素特有の緩和策例

| 攻撃 | 緩和策例 |
|--------------|--|
| 回避攻撃 | <ul style="list-style-type: none"> ➤ Adversarial Training 開発者が事前に敵対的サンプルを作成し、それを正しいラベルで学習させることで敵対的サンプルを作りにくくする緩和策 ➤ Certified Robustness あらかじめ保証したい揃動のサイズを決め、その範囲内では敵対的サンプルが存在しないことを保証する技術 ➤ Smoothing 決定境界をスムーズにさせることで敵対的サンプルを存在しにくくする |
| ポイズニング攻撃 | <ul style="list-style-type: none"> ➤ ロバスト化 訓練の手法を工夫することでポイズニング攻撃自体を適用しにくくする |
| モデル抽出攻撃 | <ul style="list-style-type: none"> ➤ 差分プライバシー 出力を生成するアルゴリズムを攪乱することで個々のデータへの攻撃を困難にする |
| モデルインバージョン攻撃 | <ul style="list-style-type: none"> ➤ 差分プライバシー 出力を生成するアルゴリズムを攪乱することで個々のデータへの攻撃を困難にする |
| メンバシップ推測攻撃 | <ul style="list-style-type: none"> ➤ 差分プライバシー 出力を生成するアルゴリズムを攪乱することで個々のデータへの攻撃を困難にする |

I-7. 検知・対処

本章では、機械学習システムの運用における攻撃の検知・対処について説明する。総務省のAI利活用ガイドライン[I-4]のAI利活用の流れにおける、「計画」や「システムの実装」で立案し、「AI構築」や「運用・利用」で実施されることを想定している。セキュリティ対策の手順における、I-5, I-6章で対策しきれなかった脅威への対応として行う。

なお、機械学習システム特有の攻撃の検知技術論文を調査し、「検知対象とする攻撃事象」と「監視するデータ」で分類・整理した結果を、付録「攻撃検知技術の概要」としてまとめている。

I-7.1. 機械学習システムセキュリティにおける検知・対処

機械学習システムにおける攻撃の検知・対処についても、一般的なサイバー攻撃対策[I-15]と同じように「検知」では攻撃の前兆や兆候を監視し、「対処」では、攻撃や被害が発覚した後の封じ込め・根絶・復旧を行う。

I-7.2. 検知

検知対象とする攻撃とその事象を決め、検知に必要なログの記録を行う。検知は、攻撃が将来起こる兆しの事象を検知する「前兆検知」と、攻撃がすでに起きたか、もしくは現在起こっていることを示す事象を検知する「兆候検知」がある。

一般的なサイバー攻撃対策においては、攻撃者の行動を分析してモデル化した MITRE ATT&CK[I-16]と呼ばれるフレームワークがあり、検知や対処に利用されている。機械学習システムへの攻撃については、同じく MITRE が公開している ATLAS[I-5]において機械学習システム特有の攻撃の戦術・手法が整理され、偵察や初期アクセスから攻撃の実行、情報窃取等の攻撃段階毎に使用される手法がまとめられており、検知対象とする事象の選定の参考になる。

I-2.2 節に挙げた 5 つの攻撃のうち、機械学習システム側において検知できる可能性のある攻撃事象の例を表 I-4 に挙げる。

表 I-4 攻撃毎の攻撃事象例

| 攻撃 | 前兆検知 | 兆候検知 |
|--------------|------------------------------------|---------------------------------|
| 回避攻撃 | | ➤ 敵対的サンプルを作成する活動 |
| ポイズニング攻撃 | ➤ 対象システムの調査 ↳ 分類ラベル数 ↳ クエリ上限 | ➤ 汚染データの訓練データへの混入 ➤ 汚染モデルの混入 |
| モデル抽出攻撃 | ➤ 攻撃用アカウントの作成 | - |
| モデルインバージョン攻撃 | | ➤ 攻撃クエリの入力 |
| メンバシップ推測攻撃 | | ➤ 攻撃クエリの入力 |

前兆検知の対象として、全攻撃に共通して、対象システムを調査する偵察活動や攻撃の下準備が考えられる。対象システムの調査は、攻撃活動の参考とするための、分類ラベル数やクエリ上限（サイズ・数）の調査等が考えられる。下準備は、実際に攻撃を実施するアカウントの作成等が考えられる。攻撃別だと、回避攻撃においては敵対的サンプルを作成する活動、ポイズニング攻撃においては、汚染したデータ・モデルをシステムに混入する活動が前兆検知の対象となりうる。なお、ポイズニング攻撃に関しては、開発プロセス（AI構築）における攻撃も想定される。そのためポイズニング攻撃の検知は、開発プロセスにおいても前兆検知が必要となる。

兆候検知では、モデルやシステムの誤動作、モデル・訓練データの窃取を実際に引き起こす活動を検知対象とする。モデルやシステムの誤動作については、回避攻撃では敵対的サンプルの入力、ポイズニング攻撃ではバックドア攻撃のトリガーを含む入力が対象となりうる。モデル・訓練データの窃取については、モデル抽出攻撃・モデルインバージョン攻撃・メンバシップ推測攻撃それぞれ、情報窃取を行うための攻撃クエリの入力が対象となりうる。

実際の検知手法・ログ設計に向けた参考情報として、機械学習システム特有の攻撃の検知技術論文を調査し、「検知対象とする攻撃事象」と「監視するデータ」で分類・整理した結果を付録「攻撃検知技術の概要」としてまとめている。

I-7.3. 対処

本節では、攻撃や被害が発生した後の対処における考え方と、I-2.2 節の 5 つの攻撃が発生した場合の対処の例について説明する。機械学習システム特有の攻撃も、一般的なサイバー攻撃と同様、すべての攻撃を未然に防ぐことは困難であるため、本節を参考に、有事の際

の応急対策が可能となる機能の設計や、調査を可能とするログ設計を行うことを推奨する。

機械学習システムセキュリティにおいても、サイバー攻撃対策と同様、攻撃や被害が発生した際には「応急対策」「調査」「恒久対策」の順で対処を行うことが推奨される。「応急対策」では発生している攻撃を一時的に緩和・停止させる処置をおこなう。次の「調査」では、システムの復旧・攻撃の再発防止のために、発生している攻撃の調査や、過去に類似攻撃がなかったかの調査を行う。最後の恒久対策では、調査で得られた情報から、攻撃の再発防止策を実施し、システムを復旧する。

各工程について説明する。

I-7.3.1. 応急対策

被害の拡大防止のため、システムの制限・停止を含む応急的な対策を行う（表 I-5）。対処として、システムの制限・停止に加え、「モデルやシステムの誤動作」を目的とする攻撃に対しては「前処理による緩和」、「モデルの窃取」「訓練データの窃取」を目的とする攻撃や回避攻撃における敵対的サンプルの生成活動に対しては「出力の偽装」が考えられる。ただし、「出力の偽装」についてはシステムの透明性や説明可能性との関係の整理が必要となる。

表 I-5 応急対策例

| 分類 | 例 |
|------------|---|
| システムの制限・停止 | <ul style="list-style-type: none"> ➢ システム停止（共通） ➢ 攻撃者アカウントの停止（共通） ➢ 入力数の制限 <ul style="list-style-type: none"> （回避攻撃の敵対的サンプル生成、モデル抽出攻撃、モデルインバージョン攻撃、メンバシップ推測攻撃） ➢ 過去モデルへのロールバック（ポイズニング攻撃） <ul style="list-style-type: none"> ◆ 誤動作した入力を正しく判定できていた過去モデル（汚染前のモデル）にロールバック ➢ 攻撃対象となったラベルの出力を制限（回避攻撃・ポイズニング攻撃） |
| 前処理による緩和 | <ul style="list-style-type: none"> ➢ ノイズ除去等による撮動緩和（回避攻撃） ➢ トリガーの除去（ポイズニング攻撃のバックドア攻撃） |
| 出力の偽装 | <ul style="list-style-type: none"> ➢ 出力ラベルの偽装 <ul style="list-style-type: none"> （回避攻撃の敵対的サンプル生成、モデル抽出攻撃、モデルインバージョン攻撃、メンバシップ推測攻撃） ➢ 確信度の偽装 <ul style="list-style-type: none"> （回避攻撃の敵対的サンプル生成、モデル抽出攻撃、モデルインバージョン攻撃、メンバシップ推測攻撃） |

I-7.3.2. 調査

システムの復旧・攻撃の再発防止のため、発生している攻撃の調査を行う。基本的には発生している攻撃の目的・手法・被害の調査、過去に類似の攻撃が行われていないかの調査、他の攻撃との組み合わされていないかの調査、となる。

表 I-6 攻撃調査の例

| 攻撃 | 攻撃の調査 | 過去の類似攻撃調査 | 組み合わせ調査 |
|--------------|---|--|---|
| 回避攻撃 | <ul style="list-style-type: none"> ➤ 攻撃対象ラベルの特定 ➤ 敵対的サンプルの作成手法特定 | <ul style="list-style-type: none"> ➤ 過去の回避攻撃探索 | <ul style="list-style-type: none"> ➤ モデル抽出攻撃からの回避攻撃の可能性調査 ➤ ポイズニング攻撃からの回避攻撃の可能性調査 |
| ポイズニング攻撃 | <ul style="list-style-type: none"> ➤ 攻撃対象ラベルの特定 ➤ トリガーの特定 ➤ 混入された汚染データ・汚染モデルの特定 ➤ 混入経路の特定 ➤ 汚染データ作成手法の特定 | <ul style="list-style-type: none"> ➤ 過去のポイズニング攻撃が原因の誤判定探索 | <ul style="list-style-type: none"> ➤ モデル抽出攻撃からのポイズニング攻撃の可能性調査 |
| モデル抽出攻撃 | <ul style="list-style-type: none"> ➤ どの程度モデルが盗まれたかの特定 (入出力からのモデル構築) ➤ 攻撃手法の特定 | <ul style="list-style-type: none"> ➤ 過去の攻撃試行の探索 | - |
| モデルインバージョン攻撃 | <ul style="list-style-type: none"> ➤ 漏えいした訓練データの特定 ➤ 攻撃手法の特定 | <ul style="list-style-type: none"> ➤ 過去の攻撃試行の探索 | <ul style="list-style-type: none"> ➤ モデル抽出攻撃からのモデルインバージョン攻撃の可能性調査 |
| メンバシップ推測攻撃 | <ul style="list-style-type: none"> ➤ 漏えいした情報の特定 ➤ 攻撃手法の特定 | <ul style="list-style-type: none"> ➤ 過去の攻撃試行の探索 | <ul style="list-style-type: none"> ➤ モデル抽出攻撃からのメンバシップ推測攻撃の可能性調査 |

I-7.3.3. 恒久対策

恒久対策では、攻撃に対する防御・緩和策の適用や、攻撃を検知する施策の導入を行う。調査で得られた情報を用いることで、効果的な防御・緩和策の適用が可能となる。

- 回避攻撃

- 攻撃に使用された敵対的サンプルを用いた敵対的訓練
- ポイズニング攻撃
 - 汚染データ・モデルを除去した上で再訓練

I-8. 参考文献

- [I-1] Organisation for Economic Co-operation and Development (OECD),
Principles on Artificial Intelligence.
<https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- [I-2] 統合イノベーション戦略推進会議, 人間中心の AI 社会原則.
<https://www8.cao.go.jp/cstp/aigensoku.pdf>
- [I-3] 経済産業省, 我が国の AI ガバナンスの在り方 ver1.1.
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/2021070901_report.html
- [I-4] 総務省 AI 社会ネットワーク推進会議, AI 利活用ガイドライン.
https://www.soumu.go.jp/main_content/000637097.pdf
- [I-5] MITRE, ATLAS. <https://atlas.mitre.org>
- [I-6] European Network and Information Security Agency (ENISA),
Artificial Intelligence Cybersecurity Challenges.
<https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges>
- [I-7] Microsoft Corporation, Threat Modeling AI/ML Systems and Dependencies.
<https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml>
- [I-8] Information Commissioner's Office (ICO),
AI and data protection risk mitigation and management toolkit.
<https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ai-and-data-protection-risk-mitigation-and-management-toolkit/>
- [I-9] National Institute of Standards and Technology (NIST), Draft NIST IR8269:
A Taxonomy and Terminology of Adversarial Machine Learning.
<https://csrc.nist.gov/publications/detail/nistir/8269/draft>
- [I-10] 産業技術総合研究所, 機械学習品質マネジメントガイドライン 第2版.
<https://www.digiarc.aist.go.jp/publication/aiqm/guideline-rev2.html>
- [I-11] 総務省×三井物産セキュアディレクション株式会社, AI セキュリティマトリックス.
https://www.mbsd.jp/aisec_portal/index.html
- [I-12] International Business Machines Corporation, Adversarial Robustness Toolbox.
<https://github.com/Trusted-AI/adversarial-robustness-toolbox>
- [I-13] CleverHans Lab, CleverHans.
<https://github.com/cleverhans-lab/cleverhans>
- [I-14] Microsoft Corporation, Counterfit.

機械学習システムセキュリティガイドライン Part I. 「本編」

<https://github.com/Azure/counterfit/>

[I-15] National Institute of Standards and Technology (NIST), NIST SP800-61 Rev. 2:
Computer Security Incident Handling Guide.

<https://csrc.nist.gov/publications/detail/sp/800-61/rev-2/final>

[I-16] MITRE, MITRE ATT&CK.

<https://attack.mitre.org>

機械学習システム
セキュリティガイドライン
Part II. 「リスク分析編」

Version 1.02
2022 年 9 月 16 日

機械学習システムセキュリティガイドライン策定委員会
機械学習システムセーフティ・セキュリティワーキンググループ

日本ソフトウェア科学会 機械学習工学研究会



II-1. はじめに

本ガイドラインは、機械学習システムの開発者（AI開発者）が開発する機械学習システムについて、機械学習特有の攻撃の観点でどのようなセキュリティリスクや脆弱性があるかを、開発者自身で分析する手法に関するガイドラインである。本ガイドラインはAI開発者が参照するための参考情報として位置づける（強制力はない）。本ガイドラインにおける「AI開発者」は必ずしも機械学習セキュリティの専門知識を有する必要はなく、一般的の機械学習システム開発者を想定する。本ガイドラインは「機械学習システムセキュリティガイドライン本編」に示される一連の手順において、AI開発者による脅威分析の部分に相当し、具体的な分析手法を紹介する。別書「攻撃検知技術の概要」では、AI開発者が攻撃の検知技術を検討する際に参考となる情報を提供するのでこちらも参照されたい。なお、紹介する脅威分析の実現例は2022年3月現在の状況であり、攻撃アルゴリズムが今後進化した場合には、対応できなくなる可能性を含めて再度検討する必要がある。また、実現例は本ガイドラインの筆者が検討したものであり、これまでに発表されている全ての攻撃に対応したものではない。

II-2. 本ガイドラインで扱う機械学習システムについて

この章では本ガイドラインで対象とする機械学習システムについて整理する。

II-2.1. 機械学習システムの構成

本ガイドラインで対象とする機械学習システムは、機械学習(Machine Learning)を用いたシステムである。機械学習システムの機械学習処理部は訓練パイプラインと推論パイプラインから構成されるのが一般的であり、図 II- 1、図 II- 2 のように表すことができる。システムによっては訓練処理を外部で行い、推論パイプラインしか行わないシステムも存在する。機械学習システムの運用に先立って、訓練パイプラインにて大量の訓練データを用いて訓練処理を行い、訓練済みモデルを生成する。そして、推論パイプラインにて推論対象データと訓練済みモデルを用いて推論処理を行って推論結果を得る。機械学習システムの構成は必ずしも、図 II- 1、図 II- 2 のものとは限らないが、本ガイドラインの内容はシステムの構成に合わせて適宜読み替えて頂きたい。

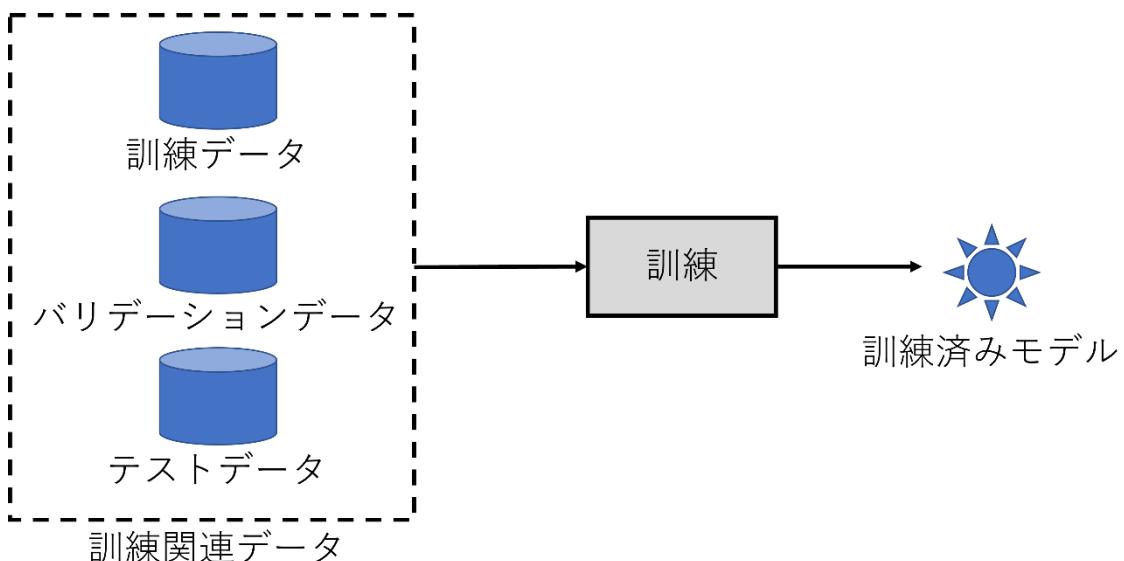


図 II- 1. 機械学習システムの訓練パイプライン

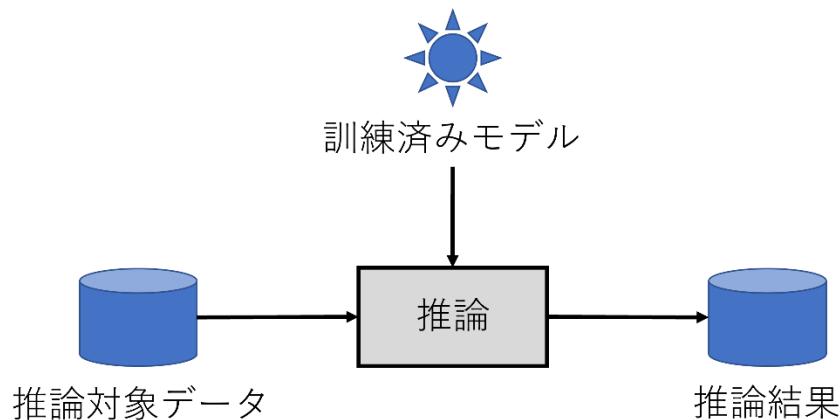


図 II- 2. 機械学習システムの推論パイプライン

II-2.2. 機械学習システムの開発プロセス

一般的な IT システムの開発と異なり、機械学習システムの開発時には、顧客からの要求に答える機械学習システムを開発するために、設計した後で試作を行い、精度や性能を評価してから正式な開発に移るケースが多い。試作の結果、期待する性能が出ていなかった場合には設計からやり直すこともある。このような試行を含んだ、機械学習システムの機械学習処理部における開発プロセスの一例を図 II- 3 に示す。この図は、機械学習システムセキュリティガイドライン本編 I-1.3.1 節で参照した総務省 AI ネットワーク社会推進会議の AI 利活用ガイドライン [II-1] における利活用の流れのうち、AI 構築部のみにフォーカスを当てて記載したものである。本ガイドラインでは、このような開発プロセスにセキュリティリスク分析のフェーズを入れることを検討する。検討結果は II-5 章で説明する。

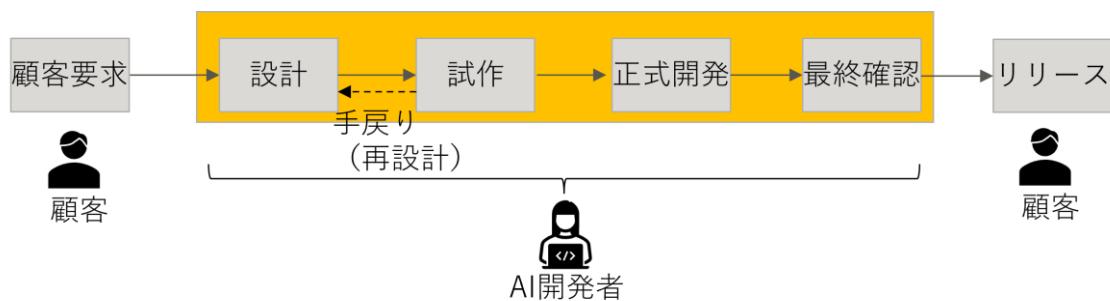


図 II- 3. 機械学習システムの機械学習処理部における一般的な開発プロセス

II-3. 機械学習システムセキュリティの概要

この章では本ガイドラインで扱う機械学習システムへの攻撃手法やその被害について整理する。

II-3.1. 機械学習への攻撃

近年、機械学習システムに対する機械学習特有の攻撃の存在が指摘されている。この攻撃は、機械学習システムに正当にアクセスしているにも関わらず、機械学習が間違えた判断をしたり、訓練データや訓練済みモデルを盗んだりしてしまうものである。攻撃者は、常に正当なアクセス権を保有して、システムを正常に操作する。システム側から見ると正常な処理であり、正当なアクセスとの区別が困難である。この点において、一般の情報セキュリティ分野における攻撃とは異なっている。(情報セキュリティ分野における攻撃では、システムに異常なデータを入力するなどしてシステムを誤動作させるものは多い)。機械学習への代表的な攻撃は機械学習システムセキュリティガイドライン本編 I-2.2 節、及び、表 II- 1 にまとめられる。

表 II- 1. 機械学習への代表的な攻撃

| | | |
|------------------|-------------------|---|
| 訓練済みモデルの判断を誤らせる | 回避攻撃 (敵対的サンプル) | 推論時に機械学習が誤判断するような推論対象データ／推論対象物を意図的に作成する |
| | ポイズニング攻撃 | 推論時に機械学習が誤判断してしまうように、訓練データに攻撃者の攻撃データを紛れ込ませて訓練させる |
| 訓練済みモデルから情報を盗み取る | モデル抽出攻撃 | 機械学習の推論処理を何回も正当に行い、攻撃者の手元に訓練済みモデルを複製する |
| | モデルインバージョン攻撃 | 機械学習の推論処理を何回も正当に行い、訓練データを攻撃者の手元で復元する |
| | メンバシップ推測攻撃 | 機械学習の推論処理を正当に行い、攻撃者の与えたデータが訓練データに含まれるかどうかを推論することで訓練データを推測する |

II-3.2. 攻撃による被害について

表 II- 1 に示した攻撃による被害を説明する。

- ・回避攻撃（敵対的サンプル）

攻撃者によって作成されたデータや物体によって機械学習システムが誤判断する。例えば、自動運転などで、カメラで撮影した道路標識からどの標識であったかを分類するよう

機械学習システムが存在した場合、道路標識に対してシステムが誤判断するように巧みに計算したテープを貼るなどの攻撃が想定される[II-2]。この標識を撮影した自動運転車は異なる標識と誤分類して事故を引き起こす。

- ・**ポイズニング攻撃**

攻撃者が機械学習システムの訓練フェーズに介入し、攻撃者が作成したデータを訓練させることができるように攻撃が成功する。これにより、機械学習システムの精度を落としたり、誤判断を起こしたりする恐れがある。また、特定のデータが入力された場合のみ誤判断を起こすような訓練をされる恐れがある。この特定のデータのことはバックドアと呼ばれる。

- ・**モデル抽出攻撃**

攻撃者が機械学習システムに何回もアクセスし、攻撃者の手元に攻撃対象のシステムを複製する。これにより、本来のサービス提供者が労力やコストをかけて訓練した機械学習システムを複製され、無料で使用される可能性がある。また、攻撃者が複製したモデルを使ってサービスを開拓する恐れもある。

- ・**モデルインバージョン攻撃**

攻撃者が機械学習システムに何回もアクセスし、攻撃者の手元で攻撃対象のシステムの訓練データを復元する。これにより、機械学習システムが訓練の際に使用した訓練データが漏洩し、プライバシーの問題を起こす恐れがある。例えば顔を分類するシステムにおいて、誰の画像を使って訓練したかが漏洩する。

- ・**メンバシップ推測攻撃**

攻撃者が機械学習システムにアクセスし、攻撃者が保持しているデータが訓練データに含まれているかどうかを推定する。これにより攻撃者に訓練データの情報が漏洩し、プライバシーの問題を起こす恐れがある。例えば既往歴を扱う機械学習システムにおいて、訓練データに特定の人物のデータが含まれているかを推定することができ、その人物に既往歴があることが分かる。

II-4. 機械学習システムを守るには

この章では II-3 章で説明した機械学習システムへの攻撃を防ぐための戦略、及び、通常の IT セキュリティの取り扱いについて説明する。

II-4.1. 機械学習システムを守る手段

機械学習システムセキュリティガイドライン本編で示されている通り、機械学習特有の攻撃からシステムを守る手段としては以下の 2 種類の手段が存在する。

1. 機械学習システムへの攻撃に対する専用手段による防御
2. 機械学習システムへの攻撃を実施困難にする運用による防御

上記の内、専用手段による防御とは、II-3.1 節で説明した機械学習システム特有の攻撃への専用防御手段のことである。現在幅広く研究され、機械学習システムセキュリティガイドライン本編 I-6 章に記載されるように多くの手法が提案されているが、どの防御手段で守っているかを知っている攻撃者については、防御を回避する攻撃ができてしまうケースがあることも指摘されている。このため、これさえ実施すれば守れるというような確固たる手段は未だ確立されていないのが現状である。このため、専用の防御手段を適用する前に、実施できる攻撃を極力減らしておくのが好ましい。実施できる攻撃を減らす手段としてシステム仕様を適切に設定したり、運用で防御したりする手段がある。例えば敵対的サンプルを生成する攻撃では、攻撃者が何回も推論処理にアクセスすることで攻撃を実現する。そこで、一定期間に推論処理にアクセスできる回数を制限するなどの対応を考えられる。このような防御は、システムに実施できる攻撃が何であるかを知り、攻撃者がその攻撃を実施するために必要な実施条件（前述の例においては、「攻撃者が推論処理へのアクセスを大量に実行でき」、かつ、「攻撃者が推論結果入手でき」、かつ、「攻撃者が機械学習システムデータ入手できる」など）を満たさなくなるようなシステム仕様を採用することで実現する。このため、実施可能な攻撃と、その攻撃の実施可能条件を知ることが重要となる。これらを知るための手段として脅威分析が重要である。

II-4.2. 通常の IT セキュリティとの関係

機械学習システムセキュリティガイドライン本編 I-1.3.3 節で示した通り、機械学習システムには機械学習システム特有の攻撃以外にも、通常の IT セキュリティ分野の攻撃が実施できる可能性がある。例えば、システムに侵入して機械学習モデルを直接盗んだりする攻撃も想定される。機械学習システムを安全にするには、このような通常の IT セキュリティ分野の攻撃に対する脆弱性と、本ガイドラインで説明した機械学習特有の攻撃の両方から守る必要がある。このうち本ガイドラインでは、機械学習特有の攻撃のみ説明する。

II-5. 機械学習システム開発プロセスにおけるリスク分析

この章では機械学習セキュリティに対策を行うための開発プロセス、及び、その問題点を整理し、目指すべき開発プロセスについて説明する。

II-5.1. 機械学習システム特有の攻撃に対するセキュリティを考慮した開発プロセス

II-4 章で説明した通り、機械学習システム特有の攻撃からシステムを守るには、リスク分析が必要である。リスク分析とは、前述の脅威分析に加えて、攻撃された際に生じる影響を分析する影響分析を含んでいる。図 II- 3 に示した通常の機械学習システムにおける機械学習処理部の開発プロセスに対してセキュリティ対応を考慮したプロセスは図 II- 4 のようになると想定される。

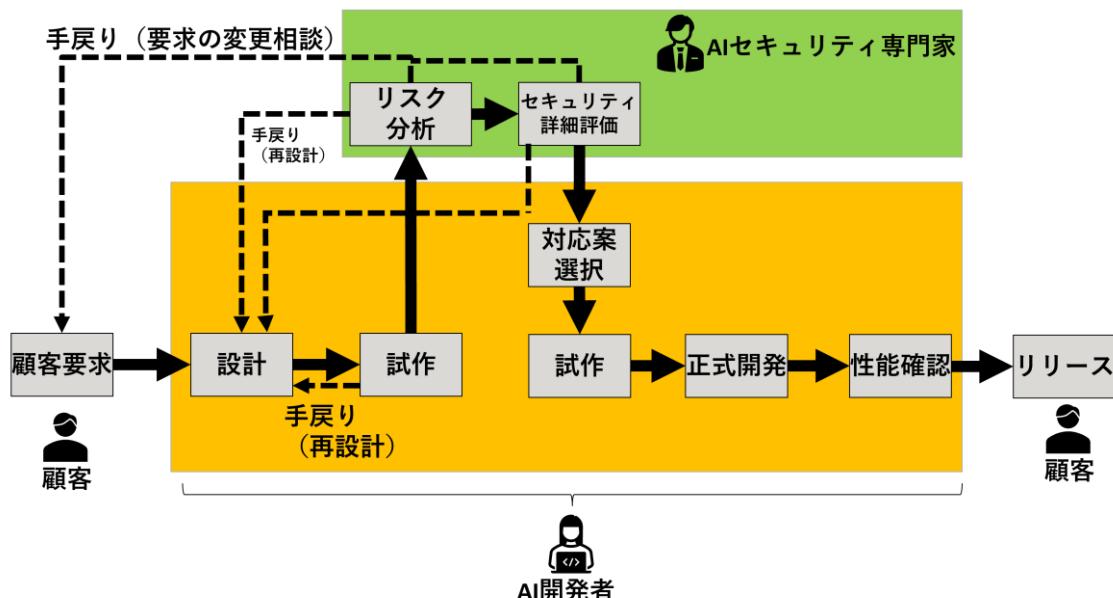


図 II- 4. 機械学習システム特有の攻撃に対するセキュリティに対応した
機械学習処理部の開発プロセス

現状、機械学習システム特有の攻撃に対するリスク分析は、機械学習セキュリティの専門家（AI セキュリティ専門家）が実施するのが一般的であると考えられる。AI セキュリティ専門家は、AI 開発者からの依頼を受けて、システムにどんなことが起きると問題となるか（どの攻撃から守りたいか、どんな被害が生じると問題となるか）などを AI 開発者や顧客からヒアリングしたのちリスク分析を行う。ここで、システムに攻撃が実施可能であるとの結論となった場合には、どのような仕様／運用にすれば守れるかの対応案を検討し、その方法を AI 開発者に通知する。通知を受けた AI 開発者はシステムを再設計し、PoC からやり直す。あるいは、顧客の要求を満足するとどうしても攻撃ができてしまうという結論となつた場合には顧客とも相談して新たに要求を作り直してから再設計する。しかし、図 II- 4 で

示した開発プロセスにおいては、AI セキュリティ専門家によるリスク分析によって多くの問題点が発見される可能性があり、リスク分析と再設計を何度も繰り返す可能性がある。このような手戻りは開発効率を下げ、開発コストの増加や納期の遅延を生じる可能性があると考えられる。このためより効率の良い開発プロセスが期待される。この問題を解決するには、現状 AI セキュリティ専門家でないと実施できないリスク分析について、AI 開発者が実施できるようにする必要がある。また、AI セキュリティ専門家そのものの人数も多くはない、AI セキュリティ専門家が各企業にいるとも限らないため、この点においても AI 開発者自身で分析を行うことが良い解決策になる。

このような AI 開発者自身が行うリスク分析を **AI 開発者向けリスク分析** と呼ぶことにする。AI 開発者向けリスク分析があれば、AI 開発者自らがリスク分析を行って、安全な仕様や運用を導くことができ、再設計を生じたとしても図 II- 4 で示したプロセスほどの多数回の手戻りは生じないと推定する。また、AI セキュリティ専門家のいない企業においてもリスク分析を実施できるようになる（ただし、脆弱性が発見された際には AI セキュリティ専門家への相談は必要となる）。AI 開発者向けリスク分析を開発プロセスに入れた例を図 II- 5 に示す。本ドキュメントでは AI 開発者向けリスク分析を実現するための手段と、著者らが実際に考案した AI 開発者向けリスク分析の一例を紹介する。なお、セキュリティ詳細評価の後で提示された対応案候補に含まれる可能性のある、攻撃検知技術については、別書「攻撃検知技術の概要」を参照されたい。

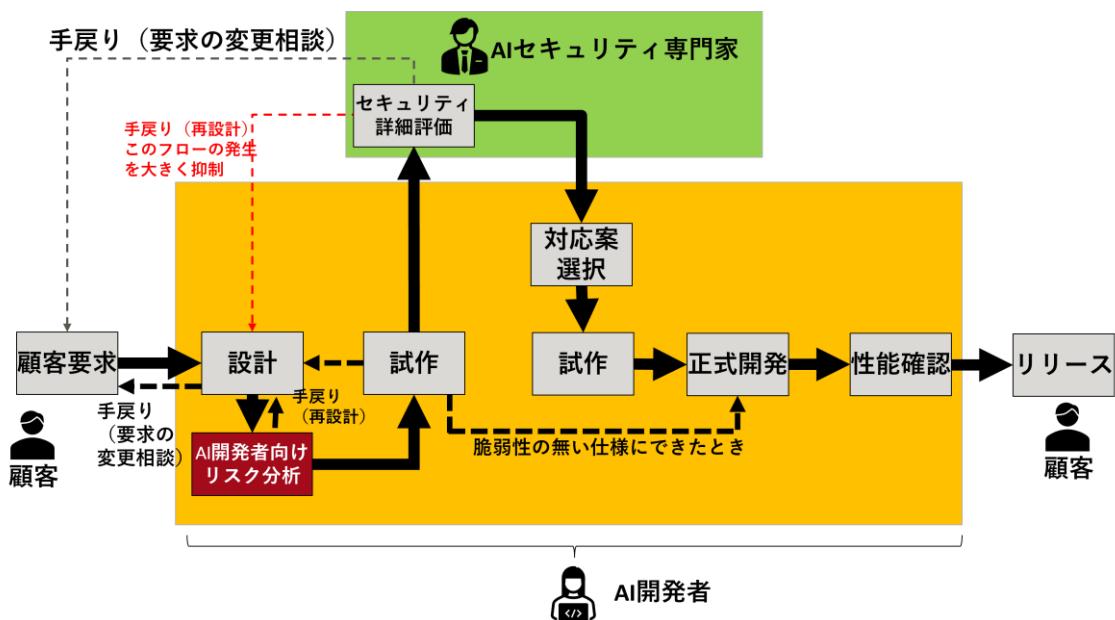


図 II- 5. セキュリティ対策による手戻りを抑制した
機械学習処理部の開発プロセス（目指すべき形）

II-5.2. 機械学習システム向けの脅威分析技術について

機械学習の脅威分析については、AI 開発者と AI セキュリティ専門家が共同で行うことができるような技術が提案されつつある。ENISA は[II-3]にて、AI システムに対する脅威や資産について、ライフサイクルを考慮しながらまとめている。また、脅威モデリングとして、資産の特定や脅威の特定、静寂性の特定などを含む 5 段階の手法を概要レベルでまとめている。マイクロソフトは[II-4]にて、AI における脅威のモデル化の考え方をまとめている。この中で、AI 開発時に確認すべき項目を質問ベースでまとめている。この質問には AI セキュリティの専門知識が必要なものも多い。このような技術は図 II- 4 や図 II- 5 のセキュリティ専門家の部分に適用可能な技術であると考えられ、AI 開発者と AI セキュリティ専門家が共同で実施する脅威分析としては参考になる。

II-6. AI 開発者向けリスク分析

この章では AI 開発者向けリスク分析の内の脅威分析部分に相当する技術として、選択回答式 AI セキュリティリスク問診技術 (AI リスク問診) を紹介する。本脅威分析は II-5 章で説明した開発プロセスで利用可能なものを目指しているが、必ずしも II-5 章の開発プロセスを前提としておらず、AI セキュリティ専門家ではない分析者、あるいは AI セキュリティ専門家を含めて、機械学習システムの分析を仕様情報から行うことができる技術である。

II-6.1. AI 開発者向けリスク分析の概要

AI 開発者向けのリスク分析では、開発中のシステムに対して、①どのような攻撃が実施できるか、また、②攻撃が実施された際にどのような被害を生じるかを洗い出す必要がある。それに加えて、③実施可能と判断された攻撃に対してどのような仕様に変更すれば、あるいはどのようなシステム運用をすれば防げるかを分析し、再設計への参考情報とする。本ガイドラインでは、①と③を解決する分析技術として、**選択回答式 AI セキュリティリスク問診 (AI リスク問診)** を紹介する。

AI リスク問診は、AI セキュリティ専門家が事前に検討した選択式質問に分析者 (AI 開発者) が回答することで分析を実施する。質問への回答後は、AI セキュリティ専門家が準備したアタックツリーが成立しているかどうかを質問への回答結果から判定する。これによりどの攻撃が実施できるかを明らかにして①を解決する。AI リスク問診では、成立したツリーを不成立にする条件が可視化されるため③も解決できる。②については、機械学習システム特有の攻撃に絞れば生じる脅威は限定されるため、どの攻撃が実施できるとどんな被害が生じるかは必ずしも AI セキュリティ専門家ではない分析者でも明らかにできる。②については機械学習システムセキュリティガイドライン本編に実施方法を含めて掲載されているので参照されたい。以降、AI リスク問診について詳しく説明する。

II-6.2. 選択回答式 AI セキュリティリスク問診 (AI リスク問診)

AI リスク問診に求められる要件は以下のとおりである。

1. 機械学習セキュリティの専門知識を持たない AI 開発者が分析できること
2. 誰が分析してもほぼ同じ結果になること
3. 分析結果の納得性が高いこと

上記要件を満たす技術として、アタックツリーを用いた分析手法[II-5]を紹介する。この技術は専門家が事前にアタックツリーを抽出しておき、抽出したツリーが成立するかどうかを分析者自らが分析する。専門家による事前準備を完了すれば、分析自体は AI セキュリティの専門知識を持たない可能性のある AI 開発者が実施することが可能である。本分析では結果がツリー形式で分かるため、結果の理解がしやすくなっている。以下、手順を詳細に説明する。筆者らが準備したアタックツリーや質問等の一例は II-7 章に記載する。それを用いて分析した例については II-8 章に記載する。

II-6.2.1. 機械学習セキュリティ専門家による事前準備手順

はじめに、機械学習セキュリティの専門家が分析のための準備をする。この準備は 1 回だけ行えばよい。

II-6.2.1.1. アタックツリーと攻撃実施可能条件の抽出

機械学習システム特有の攻撃についてのアタックツリーを構成するフェーズであり、AI セキュリティ専門家が実施する。アタックツリーは一般の IT セキュリティ分野で利用されている分析技術の一種であり、生じる脅威をトップノードとして、その脅威が発生しうる条件をツリー構造で階層的に抽出したものである。ツリー表現は自由度が高いため、一般的なセキュリティにおいてはシステム仕様が定義される前にアタックツリーを事前構成しておくことは困難である。しかし機械学習システムに限定すれば、実施可能な攻撃種や生じうる被害が限定されるため、システムの仕様が定まる前にアタックツリーを構成しておくことができる。一般的に機械学習システムへの攻撃は、一つの攻撃に対して複数の攻撃シナリオ（攻撃アルゴリズム）が存在する。ツリーを構成する際には、分析対象とする攻撃シナリオを定め、その攻撃シナリオを実施可能な条件を抽出してノードに記述する。どのシナリオを分析対象とするかは分析の粒度によって定めるが、代表的なものからツリー化していくのが好ましい。攻撃実施が可能となる条件は、論文などを参考に定める。構成したツリーの一部の例を図 II- 6 に示す。この例は回避攻撃（敵対的サンプル）に関するものである。回避攻撃（敵対的サンプル）を実施するためのシナリオを 4 つ準備している（図の左側）。いずれかのシナリオが成立すると攻撃実施可能という判断となる。図の右側は回避攻撃（敵対的サンプル）の攻撃シナリオ A1 の例である。ツリーの左側と右側が同時に成立するときに攻撃シナリオ A1 は実施可能（TRUE）となる。左側の条件は、「条件 6-2 または（条件 2-2 かつ、条件 3-1）」が成立した時に TRUE となる。右側の条件は、「条件 4-2 または条件 7-1」が成立した時に TRUE となる。このような各ノードに書かれている条件は攻撃実施可能条件と呼ばれる。

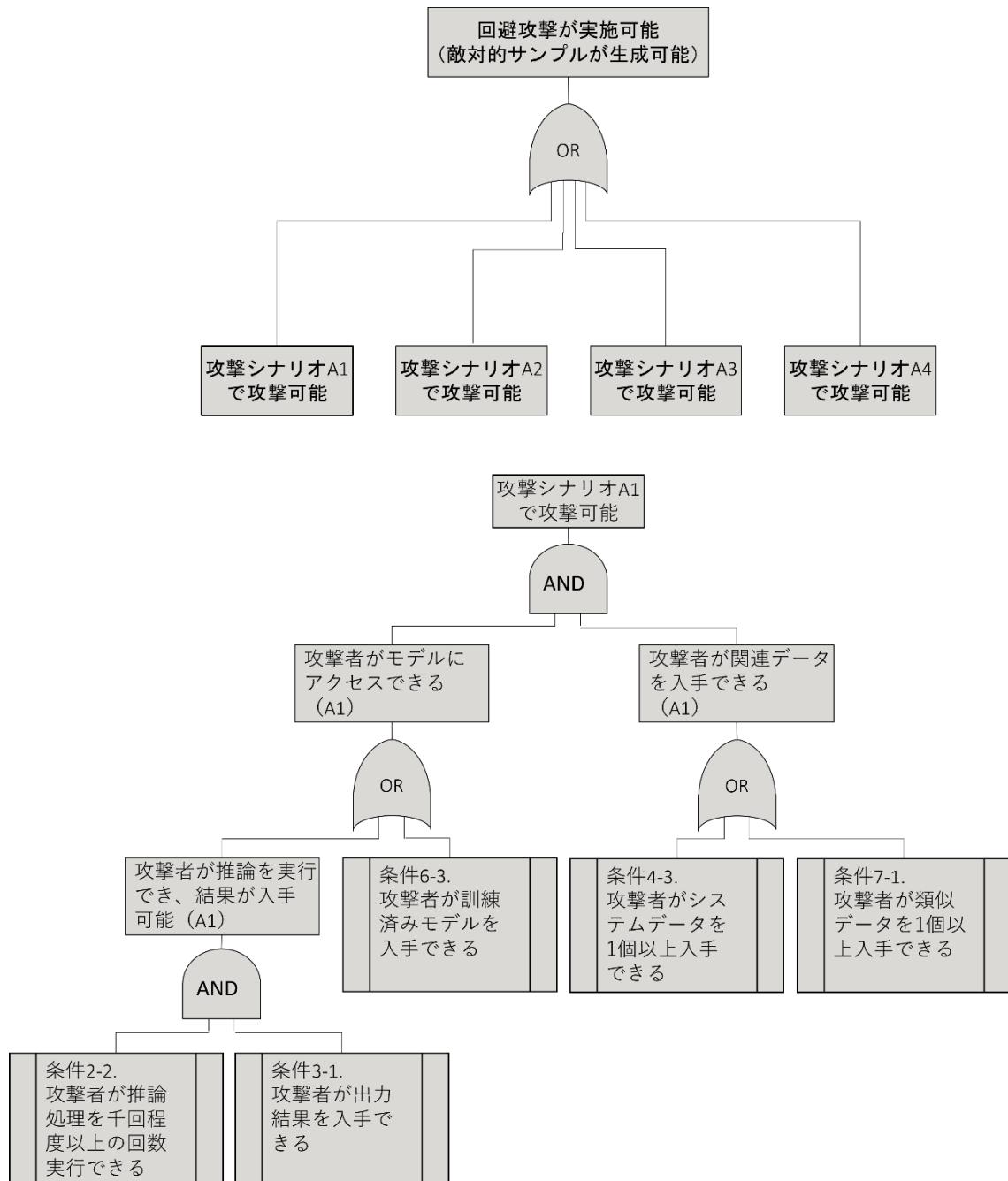


図 II- 6. 構成したアタックツリー（の一部）の例

II-6.2.1.2. 質問群の準備

アタックツリーの構成と攻撃実施可能条件群の抽出が完了したら、システム仕様が与えられたときに、そのシステムが攻撃実施可能条件を満たしているかどうかを判定するための質問群を作成する。この処理も AI セキュリティ専門家が事前に行う。分析者は AI 開発者を想定しており、必ずしも機械学習セキュリティの専門家ではないと考えられるため。質問はなるべく平易に、かつ、仕様に関する質問とした方が分析者は回答しやすい。具体例な

ども提示して分析者に理解しやすい質問を作るべきである。以下に質問の例を示す。想定攻撃者については後で説明する。

例：

質問 類似データセットの入手に関する質問

「AI システムの仕様として想定される入力データと類似のデータ（ほぼ同じ用途、ほぼ同じ種類・ジャンルのデータ）を、想定攻撃者が 1 個以上、何らかの手段で準備・入手できますか？」

- ① 類似データ 1 個とは AI が処理する最小単位のデータの集まりです。テーブルデータなら 1 行、画像なら 1 枚です。

Yes の例：顔識別 AI

想定攻撃者：利用者

一般的な顔の画像データセットを得ることができる

Yes の例：給与予測 AI

想定攻撃者：第三者

推論に使用するデータの種類（年齢、住所等）と推論対象のデータの種類（給与）が分かっていて、かつ、推論対象のデータセットとほぼ同一の分布のデータセットが入手できる場合

II-6.2.1.3. 攻撃実施可能条件の満足状況の判定用テーブルの準備

II-6.2.1.2 節で準備した質問の回答の結果を元に、II-6.2.1.1 節で抽出した攻撃実施可能条件（アタックツリーのノードに記載されている条件）を満足しているかどうかを判定するためのテーブルを準備する。例えば表 II- 2 のようなテーブルを準備する。このテーブルには、各条件が TRUE だったときに、それを FALSE にするためのシステム要件も記載しておく。FALSE にするシステム要件は対応策の検討で利用する。

表 II- 2. 攻撃実施可能条件の満足状況を判定するテーブルの例

| 条件の例 | TRUEにする条件 | 判定結果 | FALSEにするための対応策 |
|---------------------------|-------------------|-----------------------|---|
| 条件1-1 訓練処理の自由な実行が可能 | 質問1-1Aまたは1-1BがYes | TRUE or FALSE ? (埋める) | 管理者など、適かつ必要最小限の人のみ訓練処理が実行できるようにする |
| 条件2-1 推論処理を1回以上実行可能 | 質問2-1Aまたは2-1BがYes | ... | 想定攻撃者が推論処理を実行できないように設定する |
| 条件2-2 推論処理を1000回以上実行可能 | 質問2-2Aまたは2-2BがYes | | 想定攻撃者がデータ1,000個以上に対して推論処理を実行できないように設定する |
| 条件2-3 推論処理を10000回以上実行可能 | 質問2-3Aまたは2-3BがYes | | 想定攻撃者がデータ10,000個以上に対して推論処理を実行できないように設定する |
| 条件2-4 推論処理を1000000回以上実行可能 | 質問2-4Aまたは2-4BがYes | | 想定攻撃者がデータ1,000,000個以上に対して推論処理を実行できないように設定する |
| 条件3-1 推論結果入手可能 | 質問3-1または3-2がYes | | 判定結果を適かつ必要最小限の人のみに提示するようする |
| 条件3-2 確信度入手可能 | 質問3-2がYes | | 判定結果の確信度を適かつ必要最小限の人のみに提示するようにする |
| 続く | 続く | | |

II-6.2.2. 分析者による分析手順

分析者が分析を行う手順を以下に示す。この手順は II-6.2.1 節で示した準備ができていれば何回も繰り返して実施できる。

II-6.2.2.1. 分析対象システムの仕様と想定攻撃者の明確化

AI セキュリティ専門家が準備した質問に答えるための基礎となる資料として、分析者は分析対象システムの仕様を極力詳しく記述した定義資料（AI のタスク、訓練処理の実施者／実施方法／データ入力方法、推論処理の実施者／実施方法／データ入力方法、提示する出力内容／提示方法／提示先などの情報が書かれた資料）を準備する。この定義資料には、機械学習システムの出力結果を使用者に見せるか？一定の時間当たり何個のデータのクエリを許すか？システムのデータを公開するか？などの情報が含まれる。

また、分析の際に想定する攻撃者（想定攻撃者）を誰にするかも決める。分析の段階では想定攻撃者の能力を考慮して行うことになり、想定攻撃者を誰に想定するかによって分析結果に影響を受ける。従って想定攻撃者を適切に選定・設定することは極めて重要である。想定攻撃者は、システムにデータを提供する人などシステムと関連性が比較的低い人物を想定すると外部からの攻撃者の想定になり、管理者など関連性の高い人物を想定すると内部犯による攻撃者を想定することとなる。適切な想定攻撃者としては、最低限以下の人たちを想定する必要がある。

- ① AI 開発者（内部犯を想定するとき）
- ② 機械学習システムの管理者（内部犯を想定するとき）
- ③ 機械学習システムのエンドユーザ
- ④ 機械学習システムにデータが利用される人（必ずしもユーザとは限らない）

II-6.2.2.2. 質問群への回答

仕様の記述と想定攻撃者の設定が完了したら、AI セキュリティ専門家が準備した質問群に Yes/No で回答する。II-6.2.1.2 節で示したような質問群が準備されている。質問群全体の一例は II-7 章に例示する。

II-6.2.2.3. 攻撃実施可能条件の成立状況の確認

質問への回答を元に、各攻撃シナリオに相当するツリーのノード部分に記載されている攻撃実施可能条件を満たしているか (TRUE/FALSE) を判定する。これは II-6.2.1.3 節で示したような判定用のテーブルを準備し、対応する質問の回答状況から一意に定めることができる。判定用のテーブルの一例は II-7 章に例示する。

II-6.2.2.4. アタックツリーの成立状況の確認

判定用のテーブルを元に判定した攻撃実施可能条件を満たしているかの情報 (TRUE/FALSE) を、アタックツリーのノードに埋める。これにより各攻撃シナリオが成立するか、あるいは、アタックツリーそのものが成立しているかどうかが判断できる。この作業の例を II-8 章に例示する。

II-6.2.2.5. 対策の検討

成立したアタックツリーについては、想定攻撃者によってその攻撃が実施できることを示唆している。このフェーズでは想定攻撃者による攻撃を防ぐための対策を検討する。具体的には成立しているアタックツリーの構造に応じて、各ノードに記載されている攻撃実施可能条件を FALSE にするための仕様変更を検討する。検討の例を図 II- 7 に示す。この例では回避攻撃（敵対的サンプル）の攻撃シナリオ A1 が成立してしまっている。図の右側に記載されている攻撃シナリオのツリーを見ると、条件 2-2 を満たさないように機械学習システムの仕様を変更すれば、攻撃は実施しにくくなる。具体的には想定攻撃者による推論処理の実行回数を制限して一定期間に 1000 回未満にするという対応となる。ただし攻撃者が結託する可能性を考慮する場合には、全ユーザにおいて一定期間に推論処理を実行できる回数を 1000 回未満にする必要がある。一定期間とは、攻撃を防ぎたい期間であり、例えば製品の寿命までの期間などである。AI 開発者はこの条件を満たさなくするような仕様変更が実施できるかどうかを検討する。具体的にはアタックツリーのどの葉を不成立 (FALSE) にする (≒仕様変更する) にするかを検討した上で、表 II- 2 で例示した判定用テーブルの「FALSE にするための対応策」の欄に記載された対応策を参考に、葉の条件を FALSE にする仕様変更ができるかどうかを検討する。仕様変更ができないと判断した場合には、別の条件を不成立にすることを検討する。あるいは、機械学習特有の攻撃に対する専用の対策を導入することを AI セキュリティ専門家に相談する。

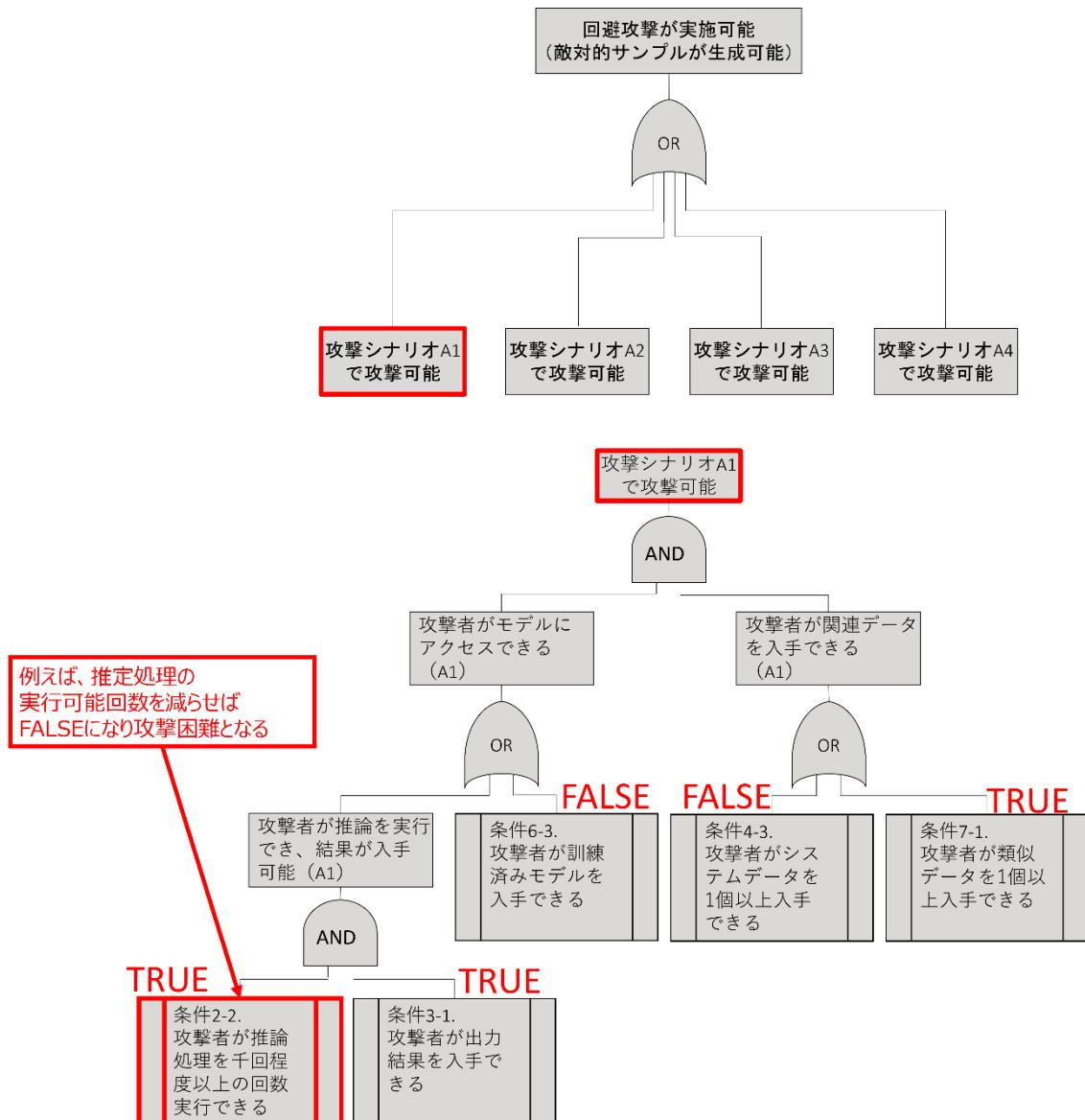


図 II- 7. 対策の検討例

II-7. AI リスク問診の実現例

この章では II-6 章で説明した AI 開発者向けリスク分析の実現例について紹介する。

II-7.1. 注意事項

[II-6]では、II-6 章で構築方法を説明した AI リスク問診についての実現例が示されている。[II-6]では、回避攻撃（敵対的サンプル）、ポイズニング攻撃、モデル抽出攻撃、モデルインバージョン攻撃のアタックツリーが掲載されている。本ガイドラインではこれに加えてメンバシップ推測攻撃についてのアタックツリーも記載するとともに、質問文等をより理解しやすいと思われる形式にしている。この実現例では各攻撃について代表的な攻撃アルゴリズムをシナリオとして定義し、アタックツリーとして抽出している。ただし 2022 年 3 月時点での実現例の一つであり、必ずしも学会等で議論されている全ての攻撃シナリオが網羅されているわけではない。今後攻撃や対策の進化により、見直される可能性／必要性があることに留意されたい。

II-7.2. アタックツリーと攻撃実施可能条件

[II-6]で記載されているアタックツリーとツリーの葉に相当する攻撃実施可能条件を以下に掲載する。各攻撃シナリオがどのような観点で構成されているかを考察した結果を表 II-3 に示す。A：回避攻撃（敵対的サンプル）、P：ポイズニング攻撃、X：モデル抽出攻撃、I：モデルインバージョン攻撃、M：メンバシップ推測攻撃である。

表 II-3. 各アタックツリーにおける攻撃シナリオの観点

| 攻撃シナリオ | アタックツリー構築の観点 |
|--------|---|
| A1 | ブラックボックス攻撃を想定した基本的な攻撃の実施条件 |
| A2 | ホワイトボックス攻撃、及び、モデル抽出攻撃よりも簡易的なモデル複製技術を利用した攻撃の実施条件 |
| A3 | モデル抽出攻撃を利用した攻撃の実施条件 |
| A4 | ポイズニング攻撃を利用した攻撃の実施条件 |
| P1 | ポイズニング攻撃の基本的な実施条件 |
| P2 | 外部や内部からのモデルを再利用した際に、モデル内部にバックドアが入っていた場合の攻撃実施条件 |
| P3 | モデル抽出攻撃を利用した攻撃の実施条件 |
| X1 | データフリー系のモデル抽出攻撃の実施条件 |
| X2 | 代表的なモデル抽出攻撃[II-7]の実施条件 |
| X3 | 代表的なモデル抽出攻撃[II-8]の実施条件 |
| X4 | 扱うデータがテーブルデータの際のモデル抽出攻撃の実施条件 |
| X5 | 扱うデータがテーブルデータ以外（画像等）の際のモデル抽出攻撃 |

| | の実施条件 |
|----|------------------------------|
| X6 | そもそもモデルを攻撃者が入手できるときの攻撃実施条件 |
| I1 | モデルインバージョン攻撃の基本的な実施条件 |
| M1 | 代表的なメンバシップ推測攻撃[II-9]のその1 |
| M2 | 代表的なメンバシップ推測攻撃[II-9]のその2 |
| M3 | 代表的なメンバシップ推測攻撃[II-9]のその3 |
| M4 | 代表的なメンバシップ推測攻撃[II-10] |
| M5 | 代表的なメンバシップ推測攻撃[II-11] |
| M6 | 代表的なメンバシップ推測攻撃[II-12] |
| M7 | 代表的なメンバシップ推測攻撃[II-13] |
| M8 | そもそも訓練データを攻撃者が入手できるときの攻撃実施条件 |

II-7.2.1. 回避攻撃（敵対的サンプル）のアタックツリーと攻撃実施好条件

回避攻撃（敵対的サンプル）についてのアタックツリーと攻撃実施可能条件は以下の通り抽出されている。

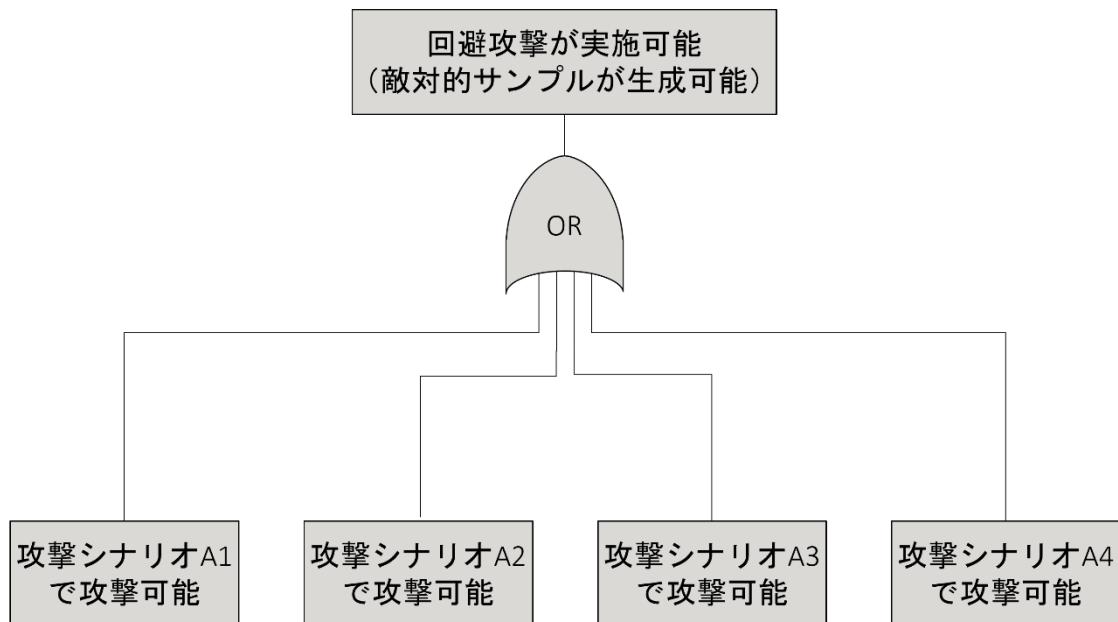


図 II- 8. 回避攻撃（敵対的サンプル）のアタックツリー（上位部分）

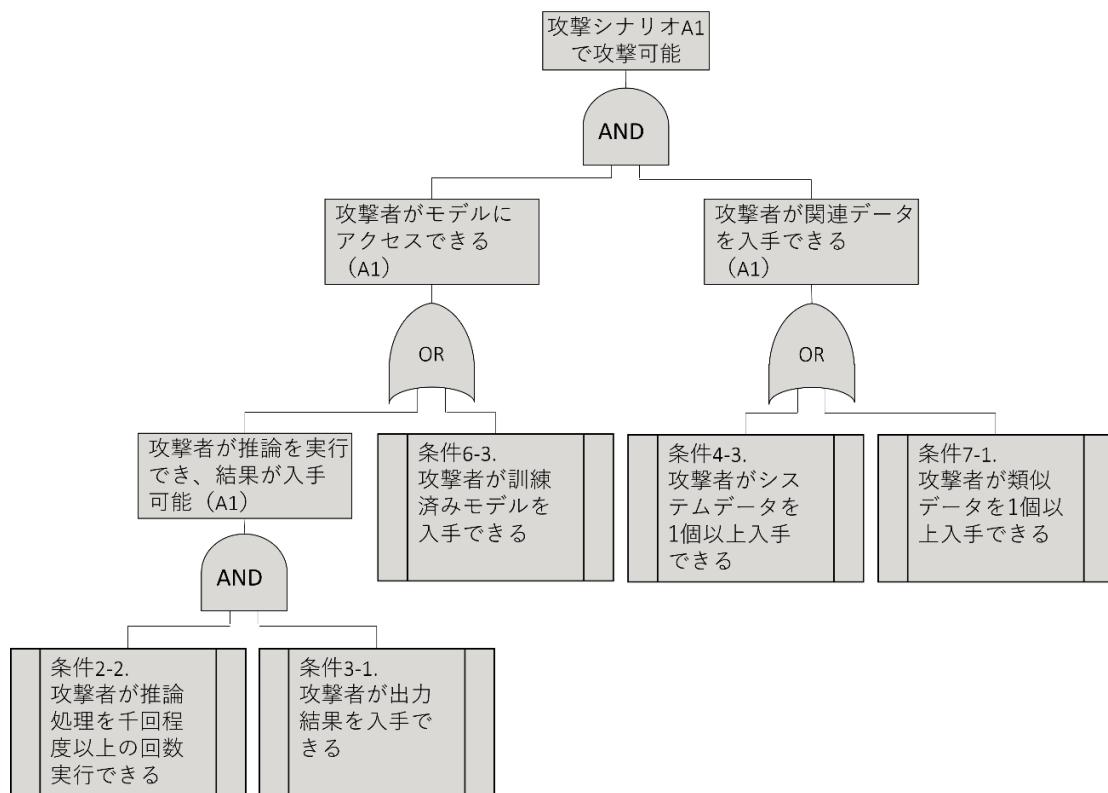


図 II- 9. 回避攻撃（敵対的サンプル）の攻撃シナリオ A1 の
アタックツリーと攻撃実施可能条件

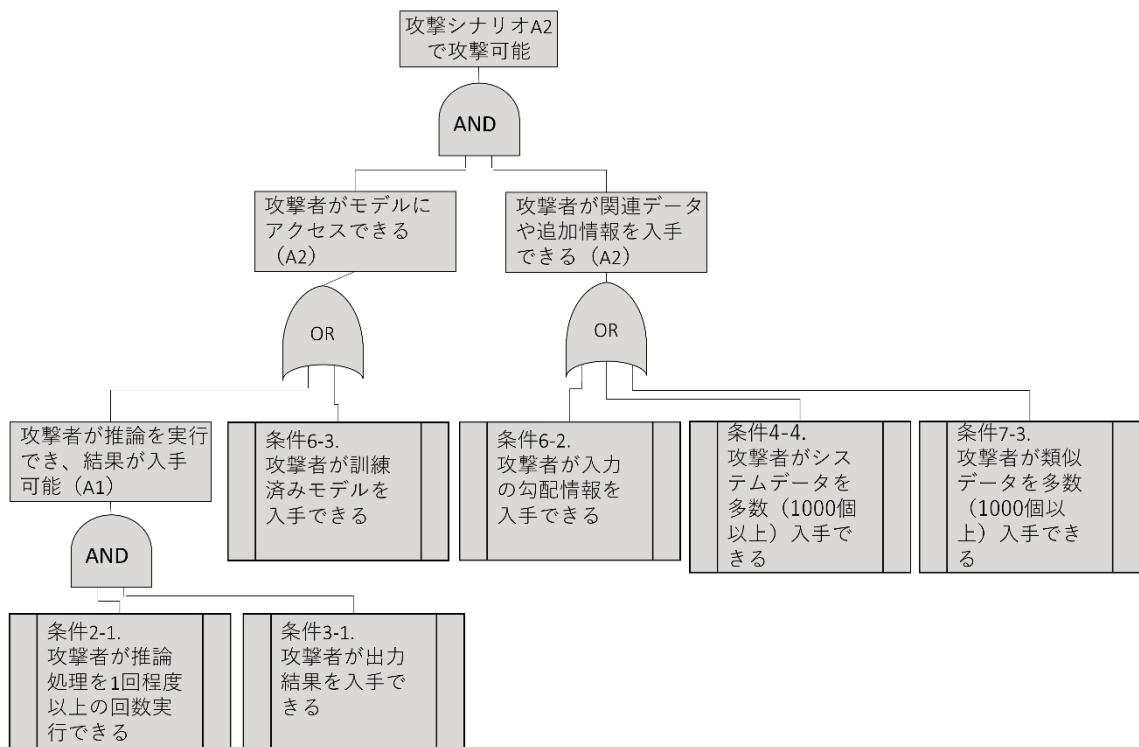


図 II- 10. 回避攻撃（敵対的サンプル）の攻撃シナリオ A2 の
アタックツリーと攻撃実施可能条件

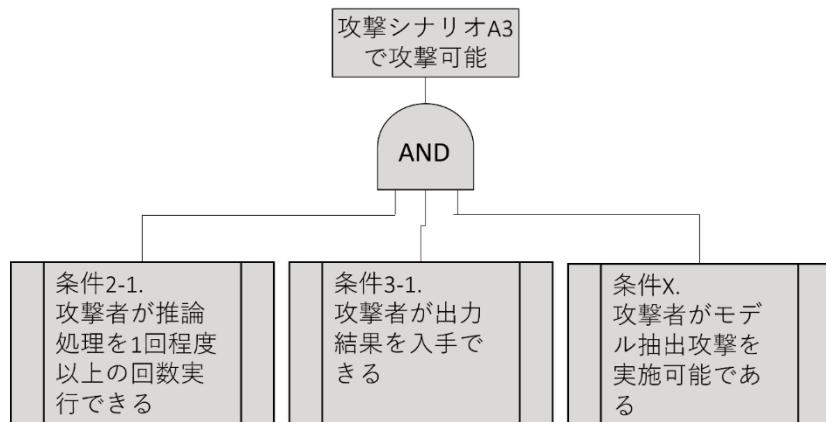


図 II- 11. 回避攻撃（敵対的サンプル）の攻撃シナリオ A3 の
アタックツリーと攻撃実施可能条件

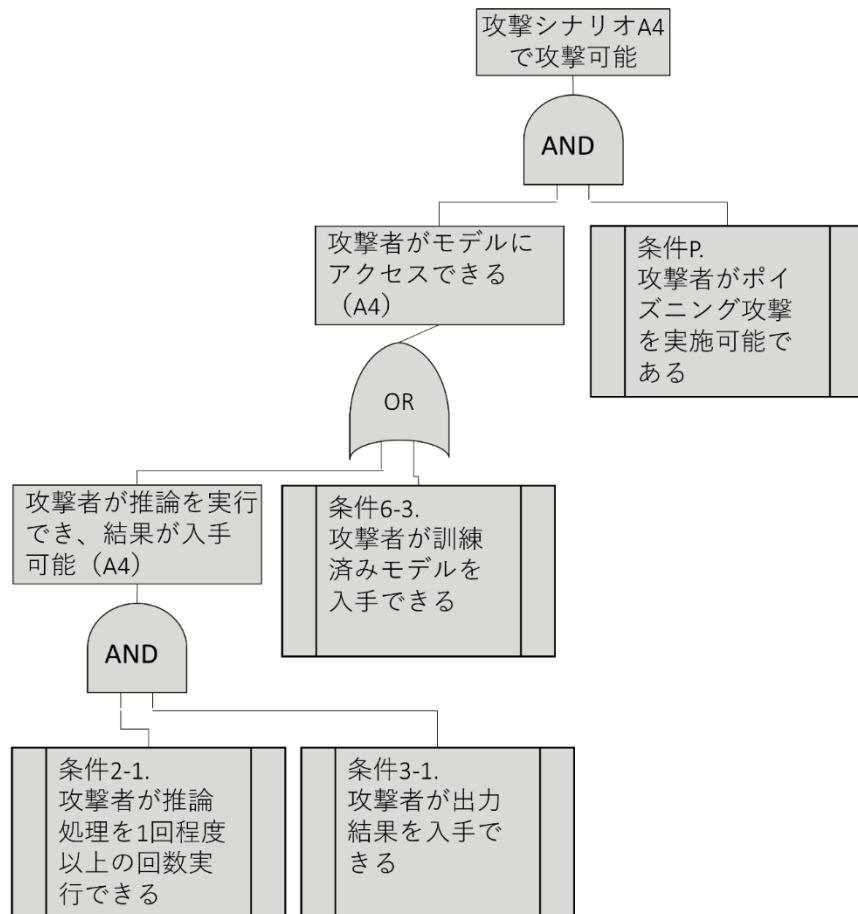


図 II- 12. 回避攻撃（敵対的サンプル）の攻撃シナリオ A4 のアタックツリーと攻撃実施可能条件

II-7.2.2. ポイズニング攻撃のアタックツリーと攻撃実施可能条件

ポイズニング攻撃についてのアタックツリーと攻撃実施可能条件は以下の通り抽出されている。

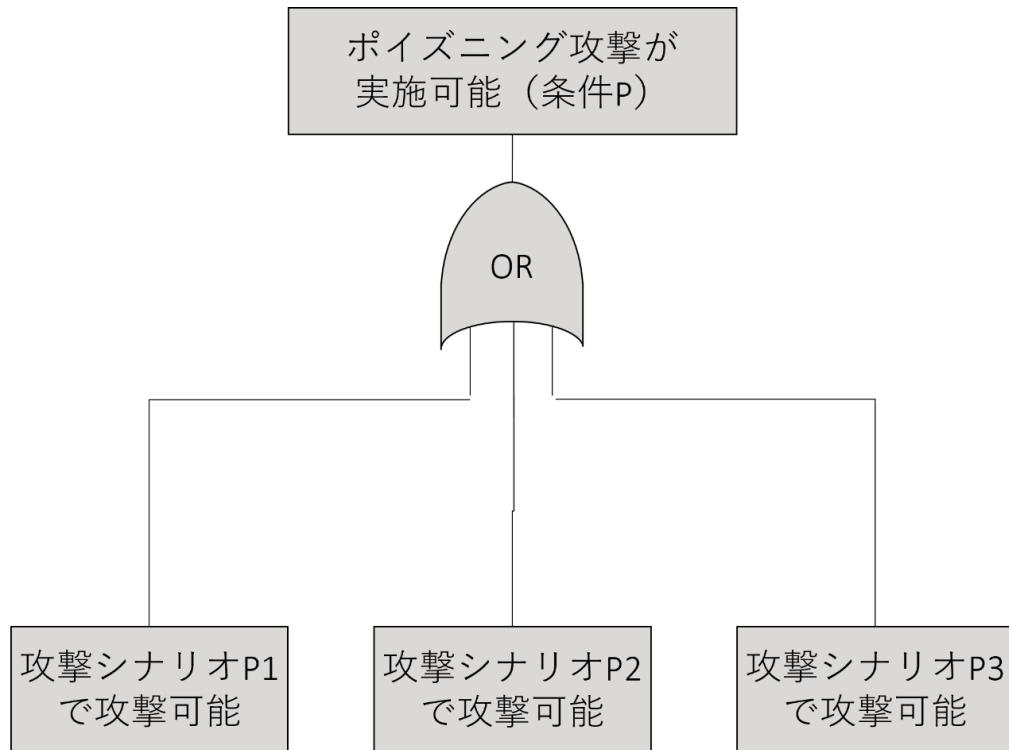


図 II- 13. ポイズニング攻撃のアタックツリー（上位部分）

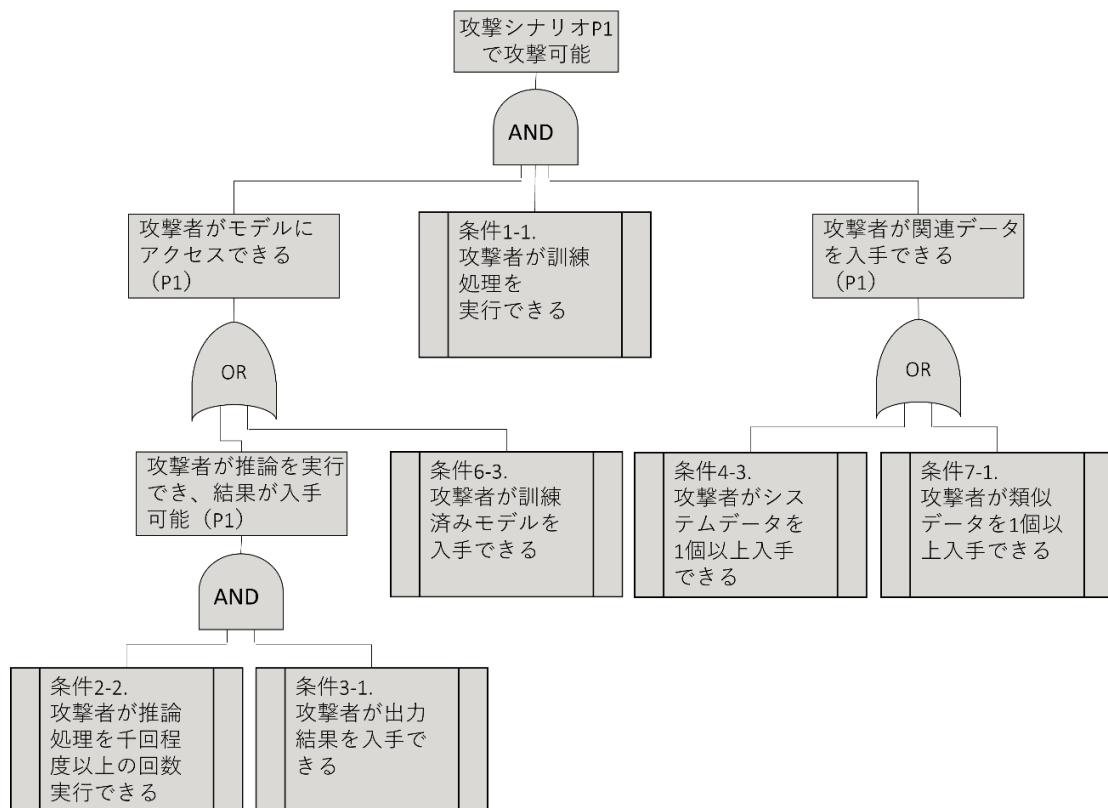


図 II- 14. ポイズニング攻撃の攻撃シナリオ P1 のアタックツリーと攻撃実施可能条件

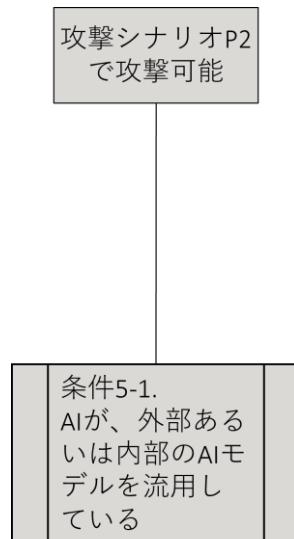


図 II- 15. ポイズニング攻撃の攻撃シナリオ P2 のアタックツリーと攻撃実施可能条件

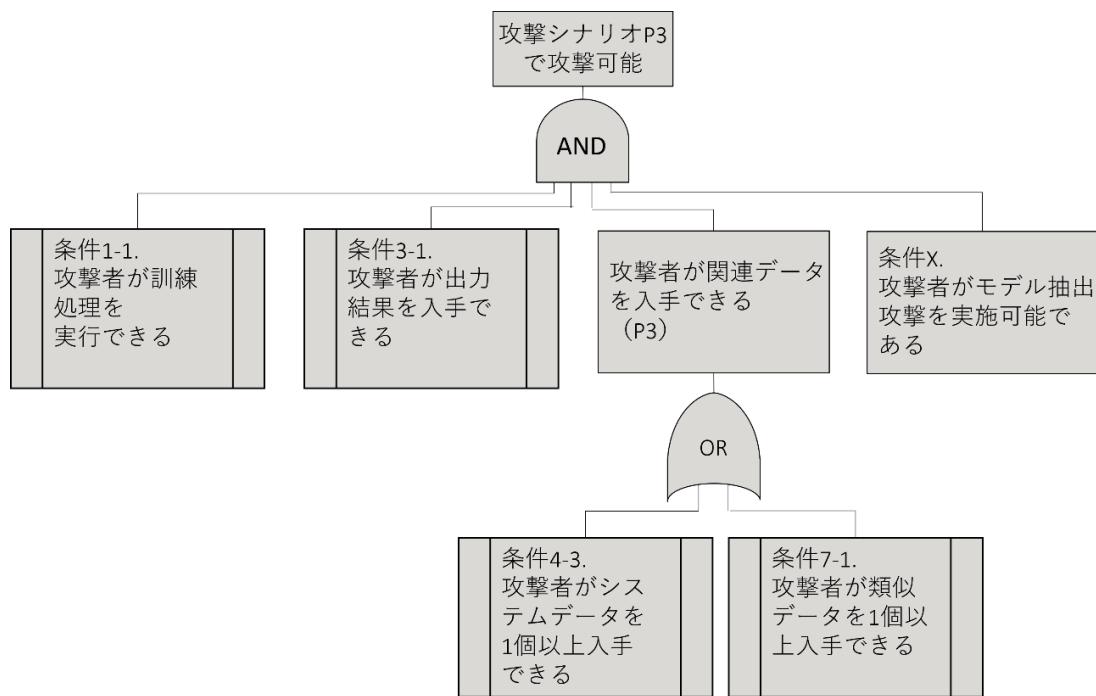


図 II- 16. ポイズニング攻撃の攻撃シナリオ P3 のアタックツリーと攻撃実施可能条件

II-7.2.3. モデル抽出攻撃のアタックツリーと攻撃実施可能条件

モデル抽出攻撃についてのアタックツリーと攻撃実施可能条件は以下の通り抽出されている。

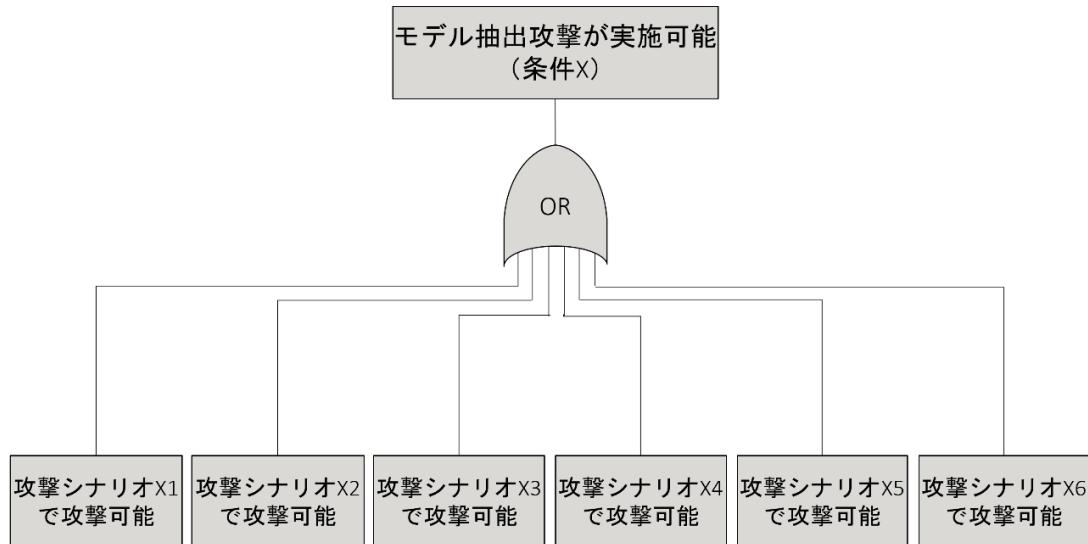


図 II- 17. モデル抽出攻撃のアタックツリー（上位部分）

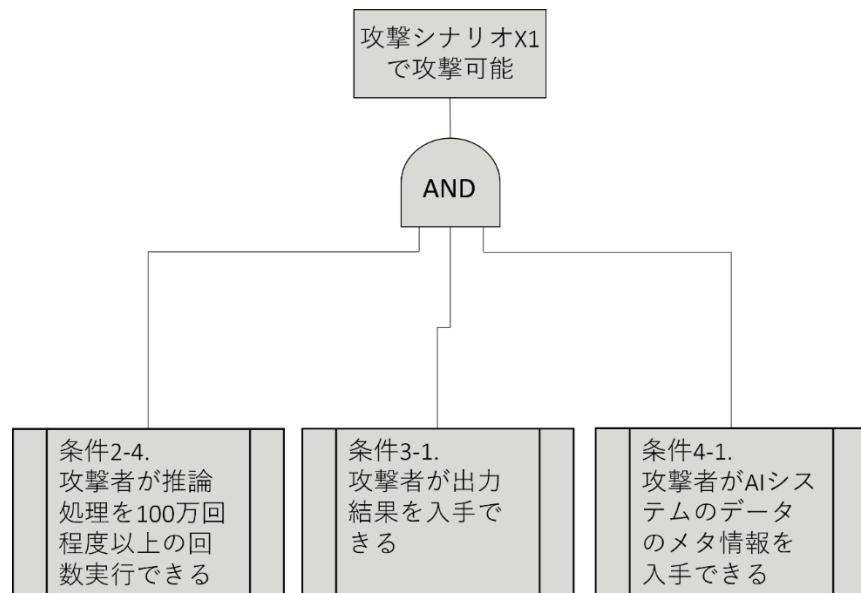


図 II- 18. モデル抽出攻撃の攻撃シナリオ X1 のアタックツリーと攻撃実施可能条件

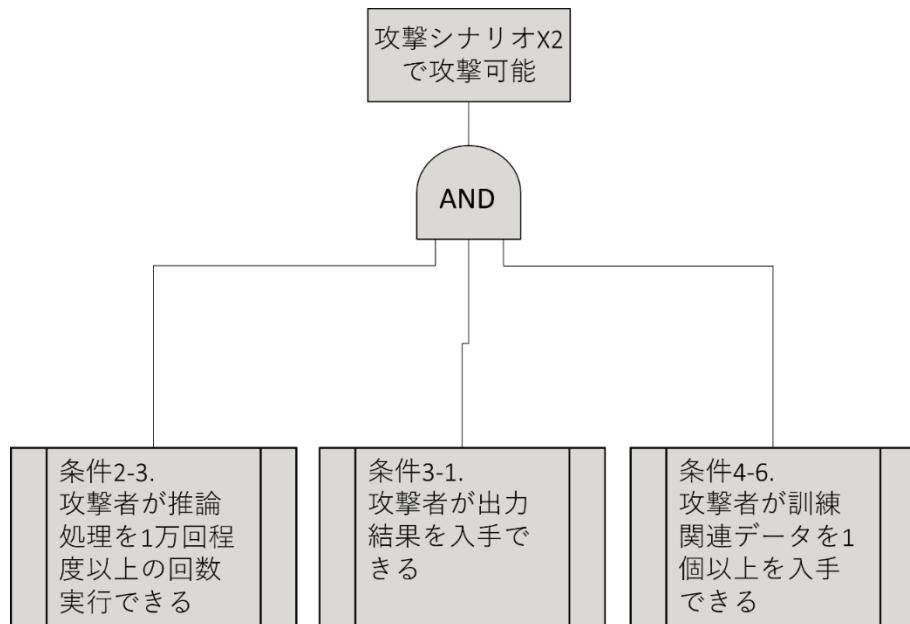


図 II- 19. モデル抽出攻撃の攻撃シナリオ X2 のアタックツリーと攻撃実施可能条件

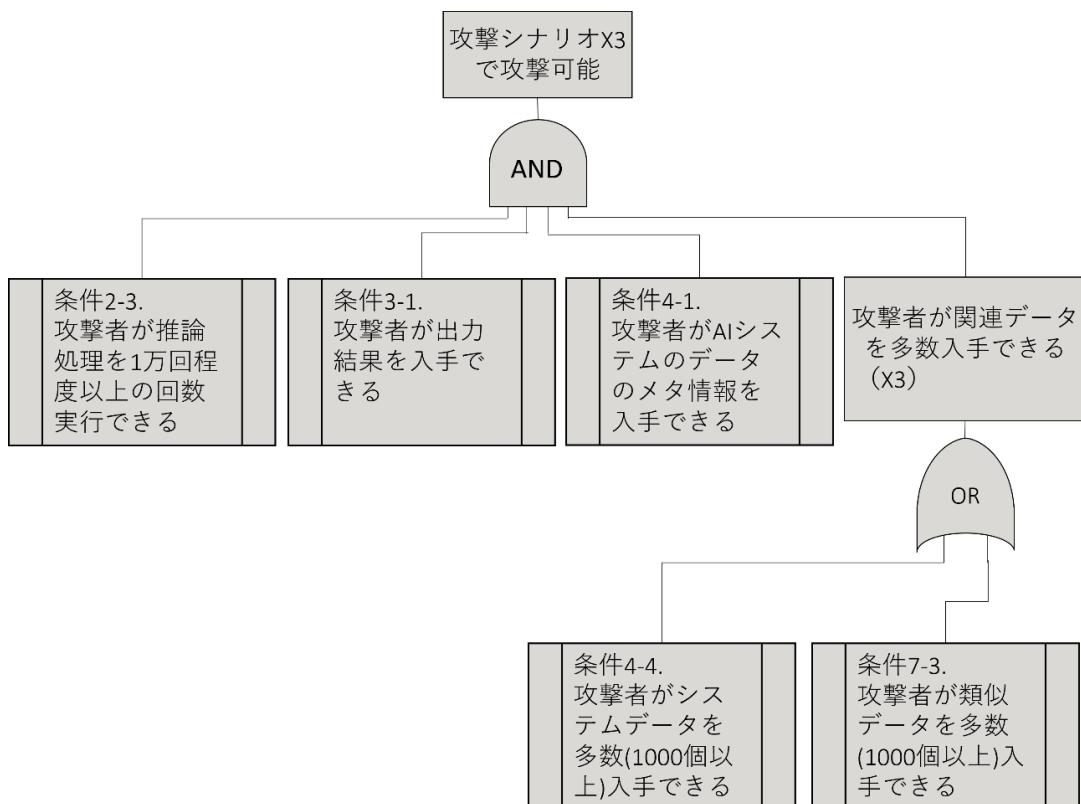


図 II- 20. モデル抽出攻撃の攻撃シナリオ X3 のアタックツリーと攻撃実施可能条件

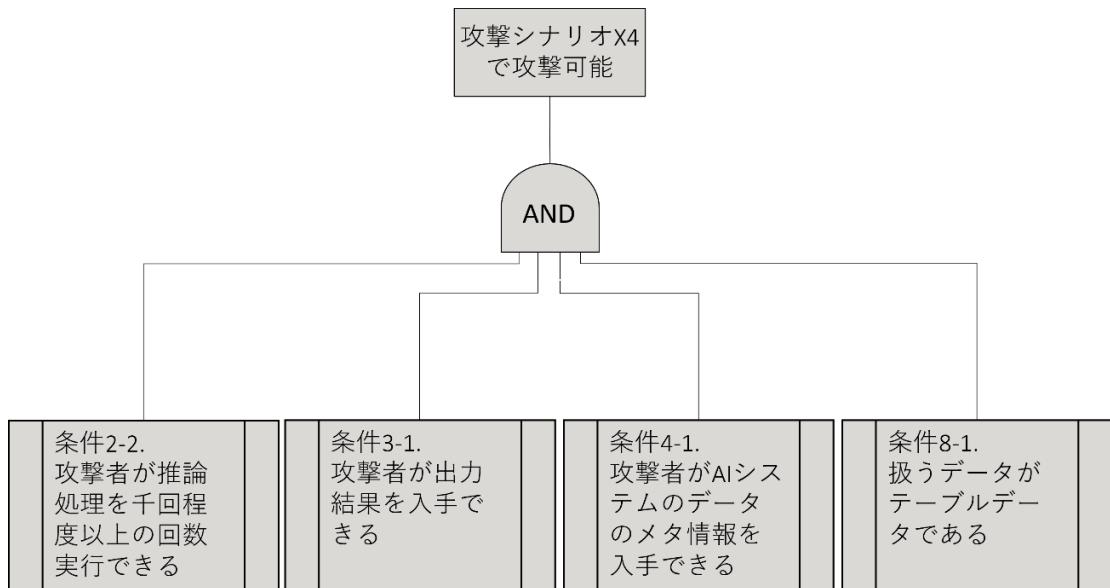


図 II- 21. モデル抽出攻撃の攻撃シナリオ X4 のアタックツリーと攻撃実施可能条件

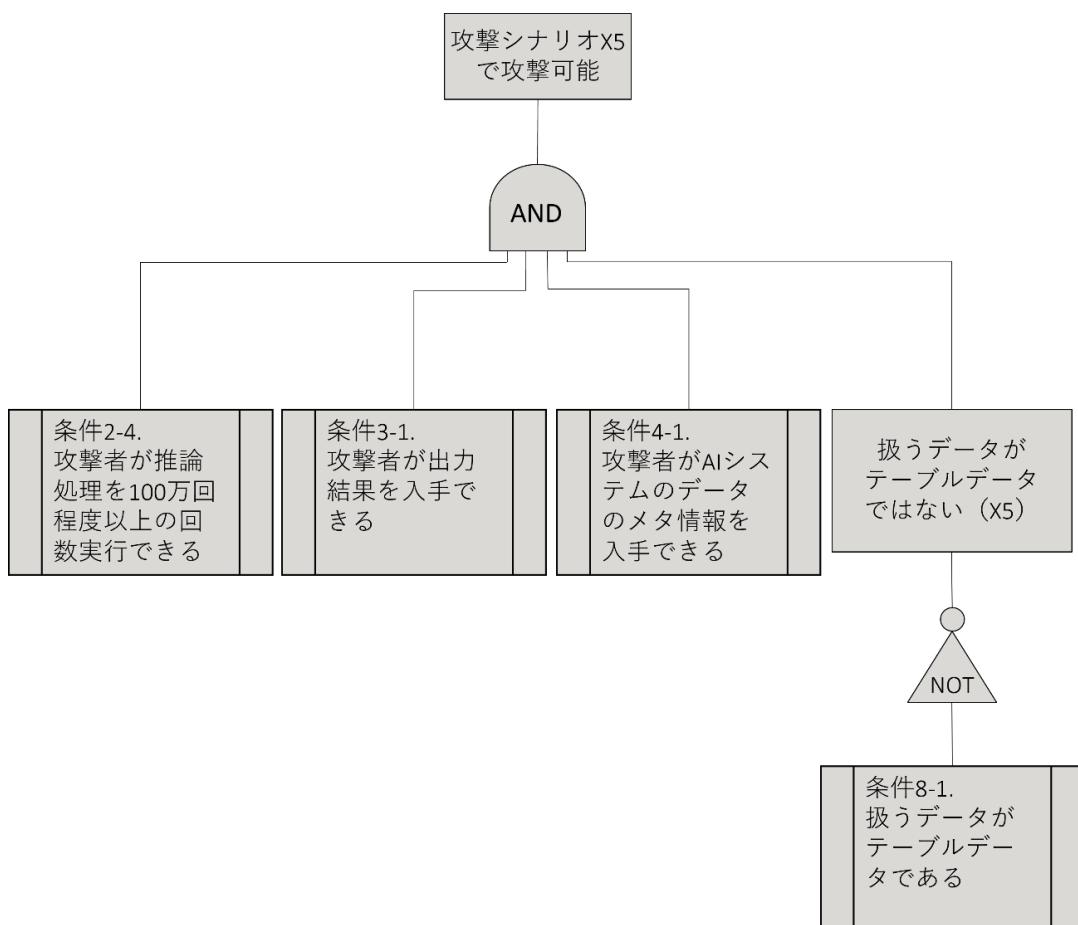


図 II- 22. モデル抽出攻撃の攻撃シナリオ X5 のアタックツリーと攻撃実施可能条件



図 II- 23. モデル抽出攻撃の攻撃シナリオ X6 のアタックツリーと攻撃実施可能条件

II-7.2.4. モデルインバージョン攻撃のアタックツリーと攻撃実施可能条件

モデルインバージョン攻撃についてのアタックツリーと攻撃実施可能条件は以下の通り抽出されている。

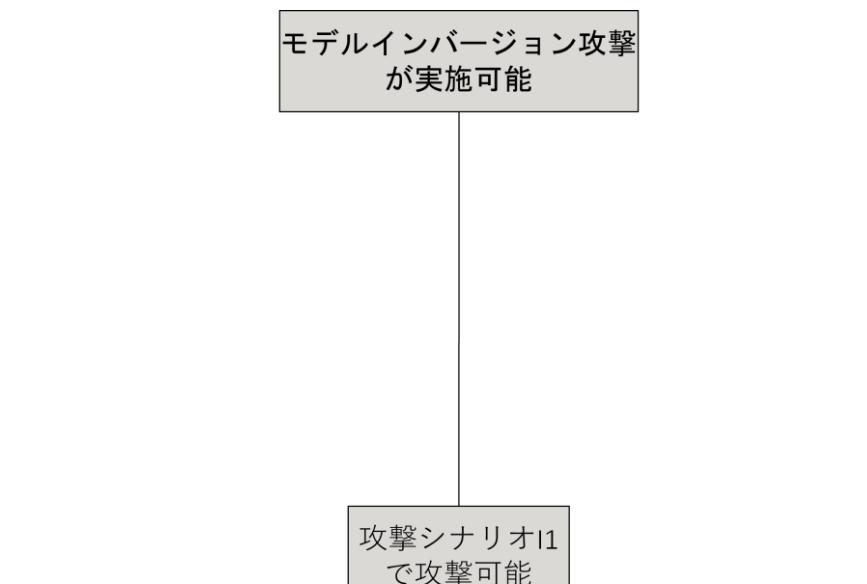


図 II- 24. モデルインバージョン攻撃のアタックツリー（上位部分）

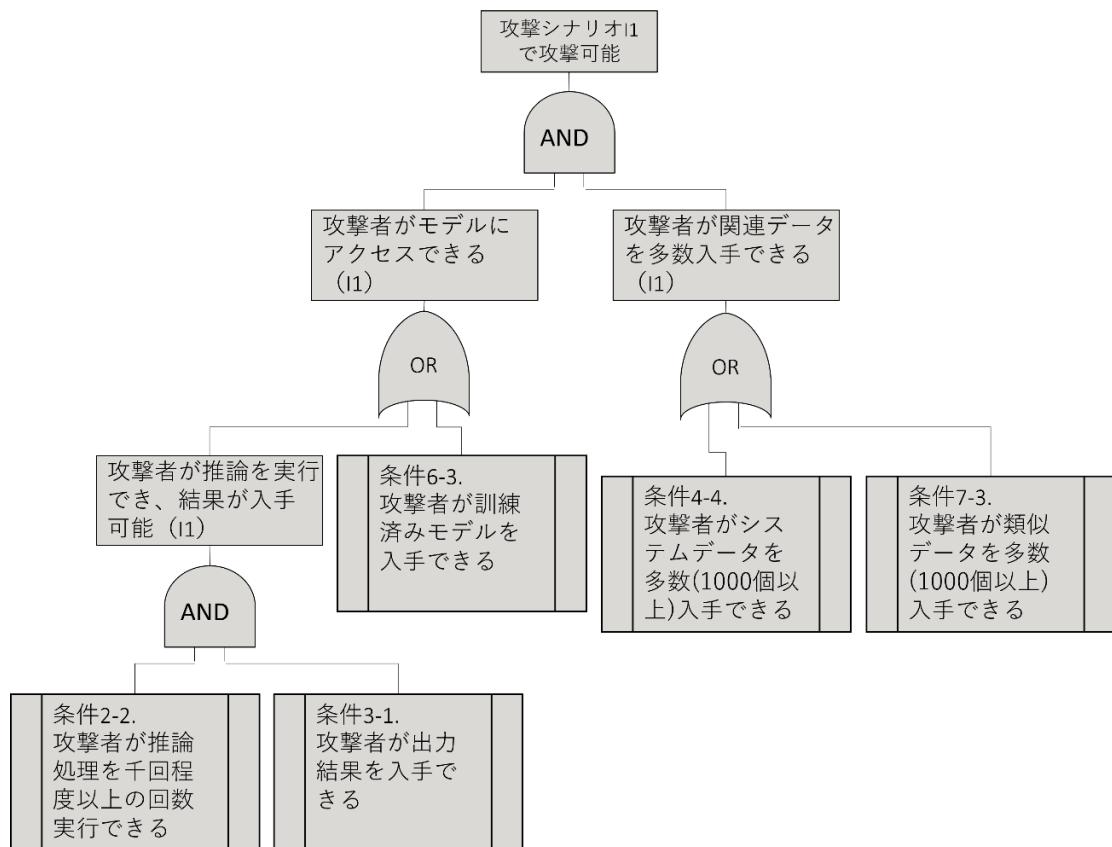


図 II- 25. モデルインバージョン攻撃の攻撃シナリオ I1 のアタックツリーと
攻撃実施可能条件

II-7.2.5. メンバシップ推測攻撃のアタックツリーと攻撃実施可能条件

メンバシップ推測攻撃についてのアタックツリーと攻撃実施可能条件は以下の通り抽出されている。

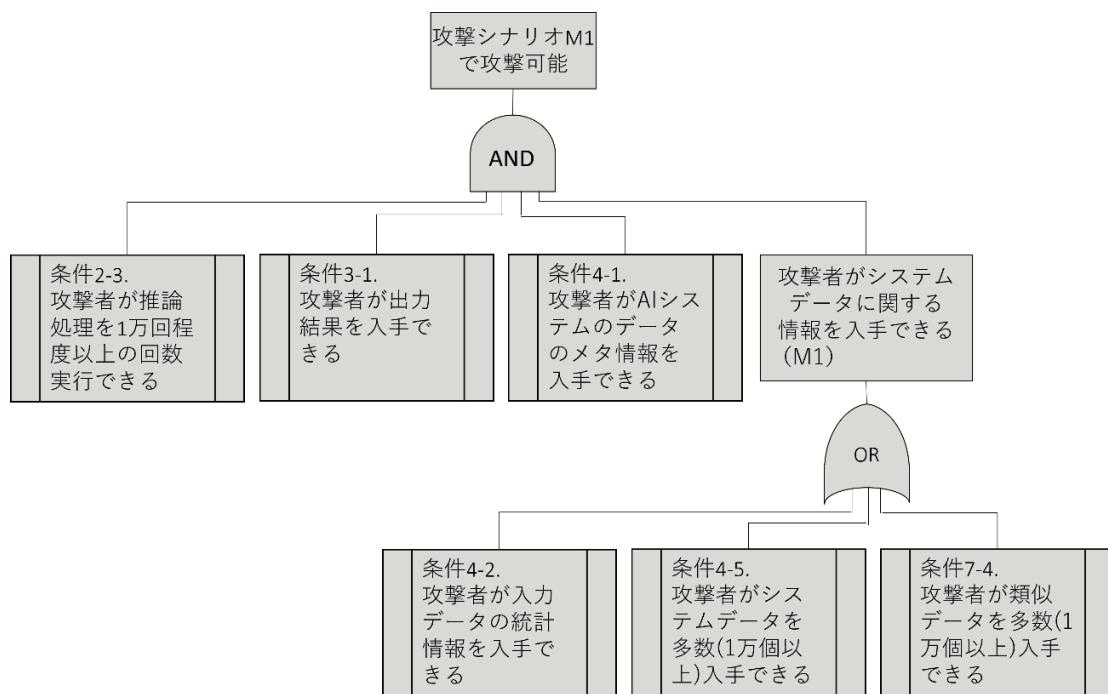
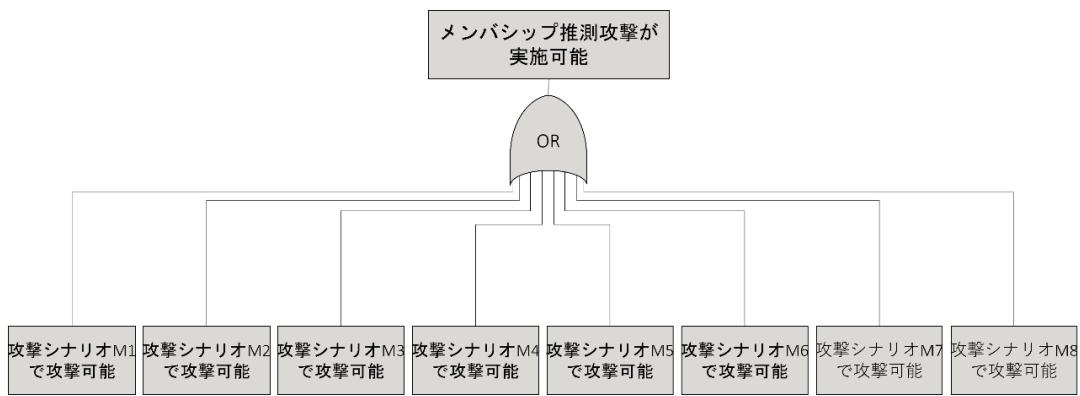


図 II- 27. メンバシップ推測攻撃の攻撃シナリオ M1 のアタックツリーと攻撃実施可能条件

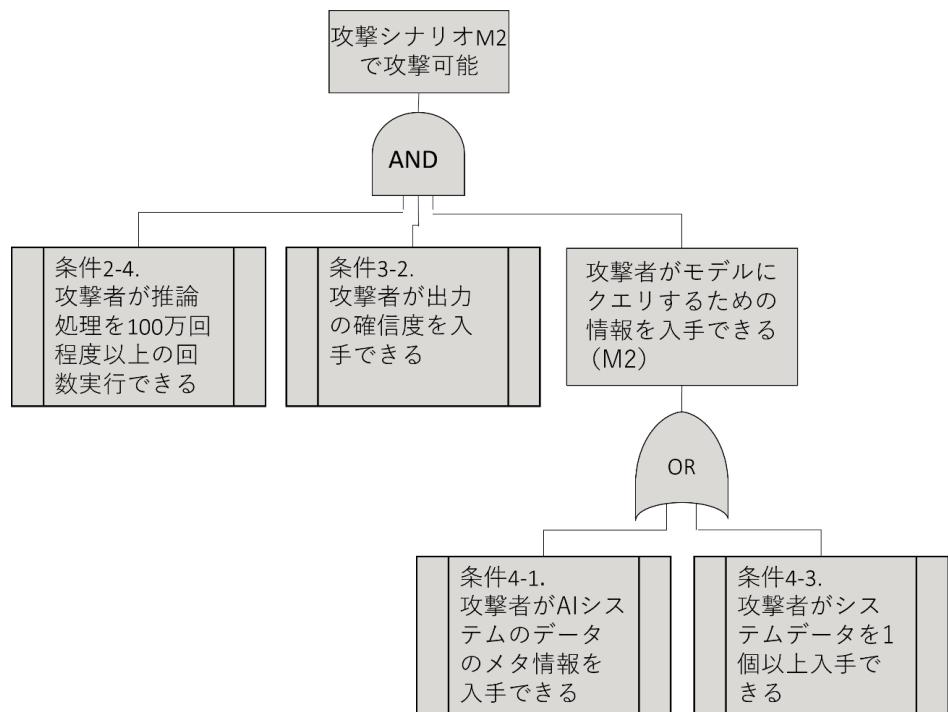


図 II- 28. メンバシップ推測攻撃の攻撃シナリオ M2 のアタックツリーと攻撃実施可能条件

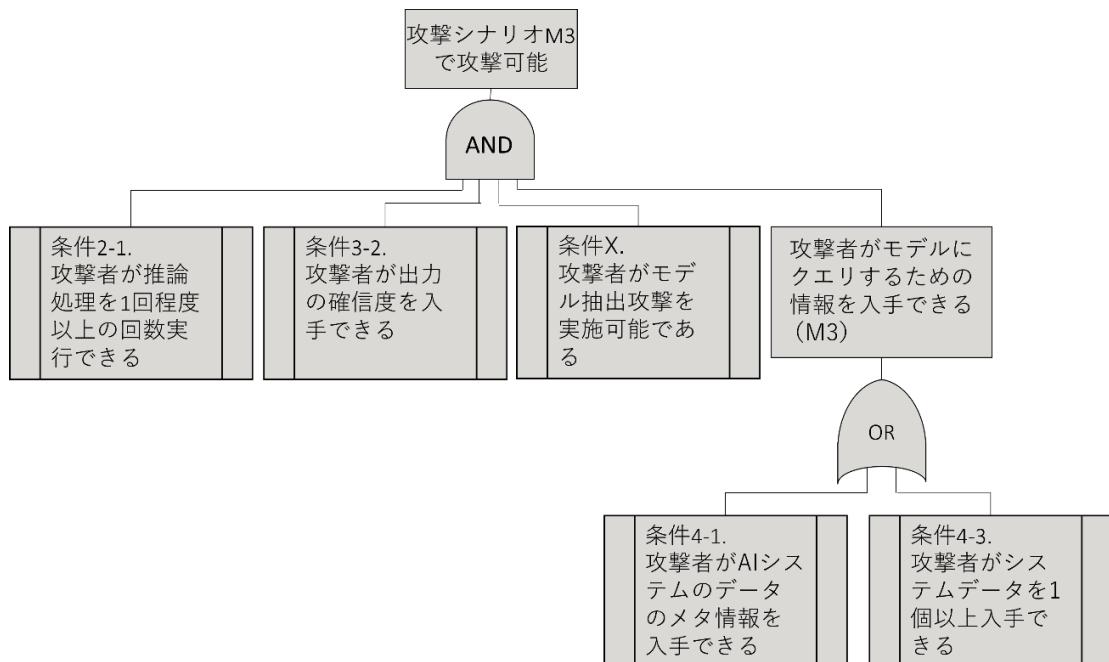


図 II- 29. メンバシップ推測攻撃の攻撃シナリオ M3 のアタックツリーと攻撃実施可能条件

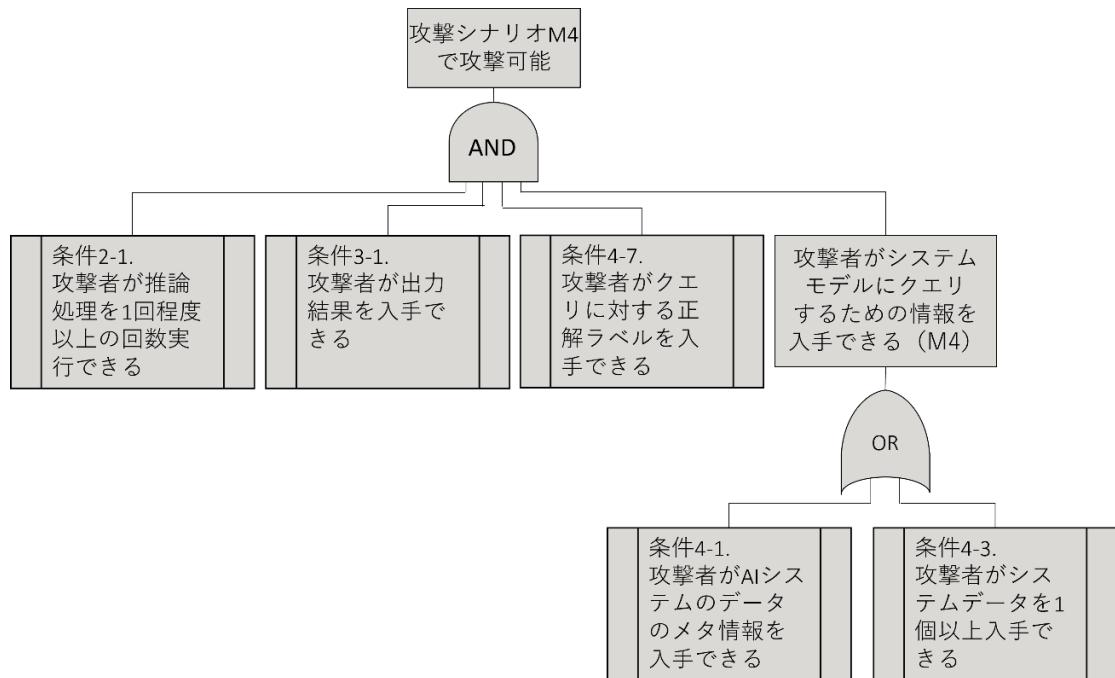


図 II- 30. メンバシップ推測攻撃の攻撃シナリオ M4 のアタックツリーと
攻撃実施可能条件

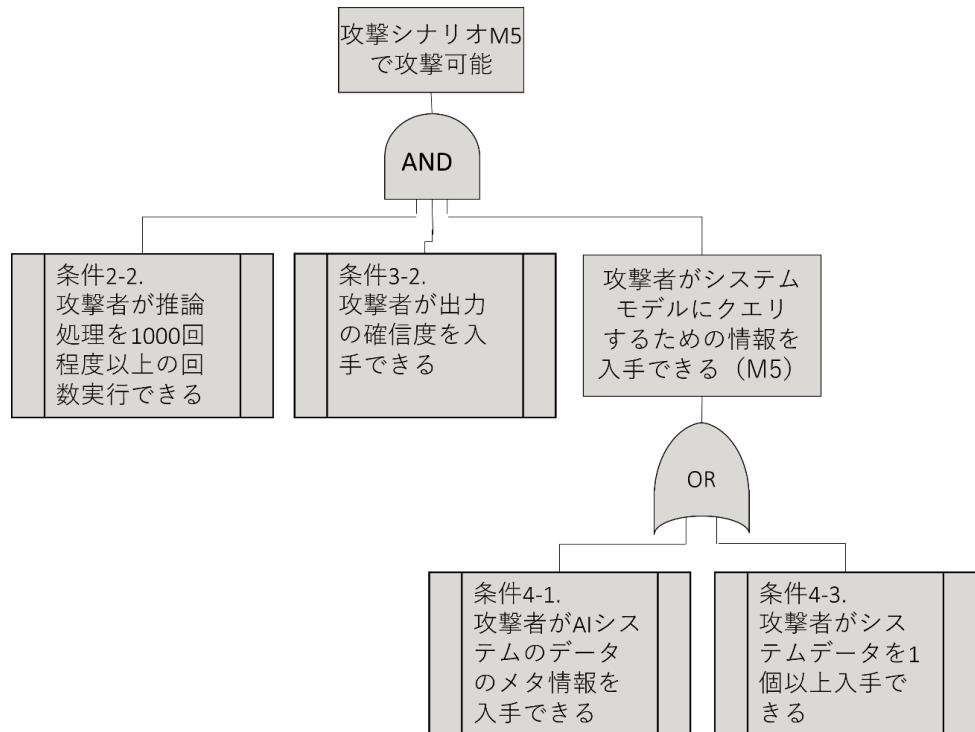


図 II- 31. メンバシップ推測攻撃の攻撃シナリオ M5 のアタックツリーと
攻撃実施可能条件

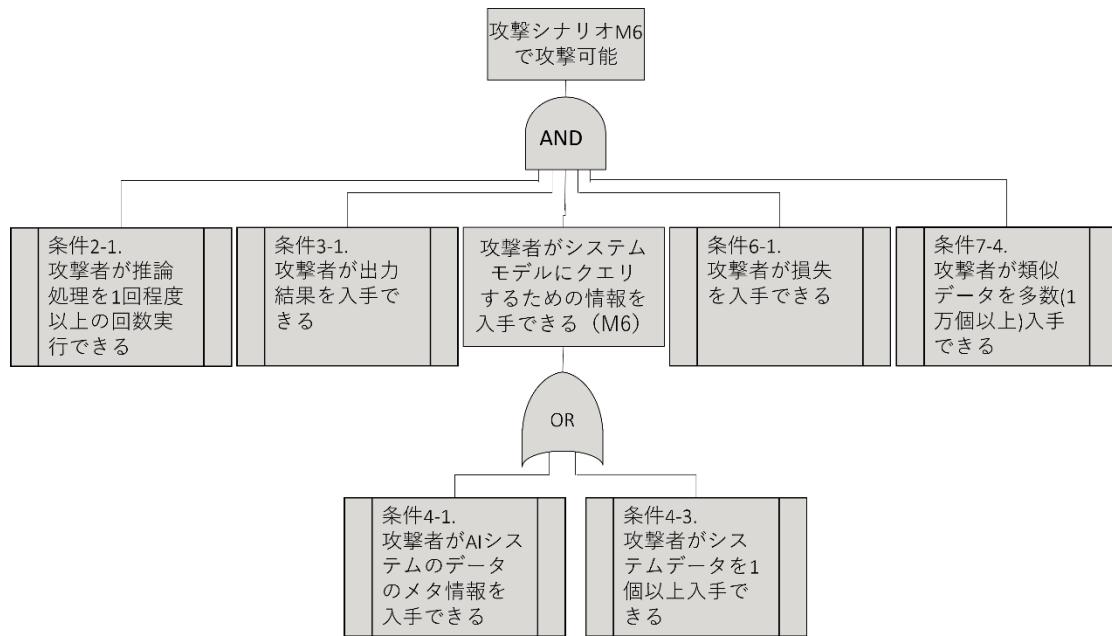


図 II- 32. メンバシップ推測攻撃の攻撃シナリオ M6 のアタックツリーと
攻撃実施可能条件

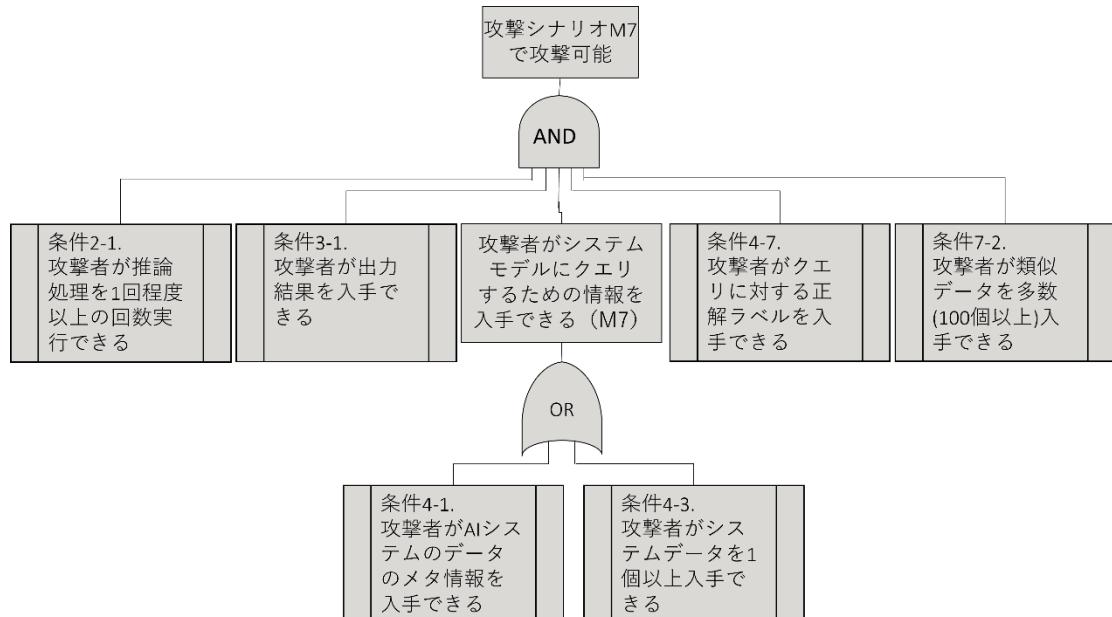


図 II- 33. メンバシップ推測攻撃の攻撃シナリオ M7 のアタックツリーと
攻撃実施可能条件

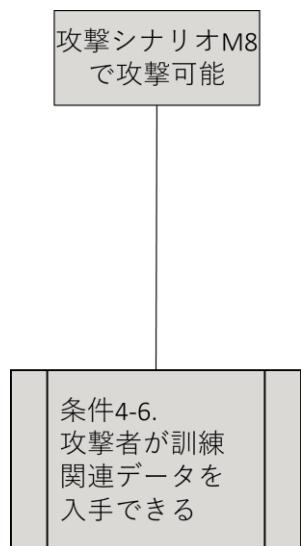


図 II- 34. メンバシップ推測攻撃の攻撃シナリオ M8 のアタックツリーと
攻撃実施可能条件

II-7.3. 質問群

II-7.2 節で掲載した攻撃実施可能条件に合致しているかどうかを聞き出すための質問についても、メンバシップ推測の判定を行うために必要な条件を追加した。さらに、分析者に理解しやすくなるように若干の改良を行った。改良後の質問は以下のとおりである。

1. 訓練処理の実行に関する質問

分析対象の AI システムが、想定攻撃者の意思で訓練処理を実行することができる場合は【1-1A】を、そうではない場合は【1-1B】をお答えください。

質問 1－1 A. 想定攻撃者の意思で訓練処理を行うことができる AI システムにおいて、自身の意図した数種類のデータで想定攻撃者が訓練処理を実行することができますか？

質問 1－1 B. 想定攻撃者でない人が訓練処理を行う AI システム、あるいは自動で訓練処理を行う AI システムにおいて、想定攻撃者が意図したデータを訓練データに混入できますか？

2. 推論処理可能なデータの個数に関する質問

分析対象の AI システムが、想定攻撃者の意思で推論処理を行える場合は【2-1A】を、自動で推論処理を行うシステムの場合は質問【2-1B】をお答えください。

質問 2－1 A. 想定攻撃者が準備した推論対象データ 1 個以上に対して推論処理を実行することができますか？

①推論対象データ 1 個とは AI が処理する最小単位のデータの集まりです。テーブルデータなら 1 行、画像なら 1 枚です。

②想定攻撃者が推論処理を実行可能なデータの個数は、運用期間や推論処理の実行間隔などを考慮して導出してください。

③想定攻撃者が複数の利用者アカウントを作成できる場合、各利用者アカウントが実行した推論処理の合計数も考慮してください。

質問 2－2 A. 【2-1A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000 個以上に対して推論処理を実行することができますか？

質問 2－3 A. 【2-2A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 10,000 個以上に対して推論処理を実行することができますか？

質問 2－4 A. 【2-3A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000,000 個以上に対して推論処理を実行することができますか？

質問 2－1 B. 想定攻撃者が推論対象データを 1 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？

- ①推論対象データ 1 個とは AI が処理する最小単位のデータの集まりです。テーブルデータなら 1 行、画像なら 1 枚です。
- ②想定攻撃者が推論処理を実行可能なデータ個数は、運用期間や推論処理の実行間隔などを考慮して導出してください。

質問 2－2B. 【2-1B】が Yes であった場合、想定攻撃者は【2-1B】において、推論対象データを 1,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？

質問 2－3B. 【2-2B】が Yes であった場合、想定攻撃者は【2-1B】において、推論対象データを 10,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？

質問 2－4B. 【2-3B】が Yes であった場合、想定攻撃者は【2-1B】において、推論対象データを 1,000,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？

3. 出力結果の入手に関する質問

質問 3－1. 想定攻撃者にモデルの判定結果を提示しますか？

あるいは、想定攻撃者は判定結果を類推することができますか？

- ①判定結果とはモデルからの出力のことで、例えば分類タスクの場合には分類ラベル、回帰などの予測 AI においては予測結果などを指します

質問 3－2. モデルの確信度を一部でも想定攻撃者に提示しますか？

4. モデルが扱うデータの入手に関する質問

質問 4－1. 推論対象データや訓練データを作成できるフォーマット情報（データ形式、サイズの情報、画像の種類、データの次元等のメタ情報）を想定攻撃者は知ることができますか？

①メタ情報とは、テーブルデータを扱うデータの場合は AI システムの入力データのフォーマット（行・列の数、及び、要素の順序など）、画像を扱う AI システムにおいては縦横のピクセル数などを指します

②推論処理の実行 API がある場合、API を実行する際の入力データのメタ情報のことを指します

質問 4－2. AI システムに用いられた訓練データの統計情報を想定攻撃者は知ることができますか？

①統計情報とは、テーブルデータを扱うデータの場合の各列の数値データの平均や分散などを指します。

質問 4－3. AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ) を想定攻撃者が、1 個以上入手できますか？

- ①入力データ 1 個とは AI が処理する最小単位のデータの集まりです。テーブルデータなら 1 行、画像なら 1 枚です。
- ②想定攻撃者が AI システムのタスクと入力データのメタ情報を知ることができます、推論対象データを生成・準備することができる場合は Yes です。
- ③推論処理の実行 API がある場合、実行 API への入力データが入手できるかを指します

質問 4-4. 【4-3】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or テストデータ or 推論対象データ)を想定攻撃者が、1,000 個程度以上入手できますか？

- ①一人の想定攻撃者は 1 個しか入手できない場合でも複数の想定攻撃者が合計で 1,000 個以上入手できる場合は当てはります

質問 4-5. 【4-4】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ)を想定攻撃者が、10,000 個程度以上入手できますか？

質問 4-6. 訓練関連データに関して、想定攻撃者が 1 個以上入手できますか？

- ①訓練関連データ 1 個とは AI が処理する最小単位のデータの集まりです。テーブルデータなら 1 行、画像なら 1 枚です。
- ②訓練関連データは、訓練データ、バリデーションデータ、もしくはテストデータのことです

質問 4-7. AI システムへ入力したデータに対する正解ラベルを想定攻撃者が入手できますか？

- ①正解ラベルとはモデルが推論したラベルではなく、真の正解ラベル(Ground Truth) です。

5. モデルの流用に関する質問

質問 5-1. 外部や内部から入手した訓練済みモデルを一部でも流用し、転移性を利用して AI を構築していますか？

- ①インターネットから入手したモデルを内部に流用している場合やインターネット以外でもあまり信頼できない入手先から入手したモデルを流用している場合に当てはります

6. モデル情報の入手に関する質問

質問 6-1. 推論処理の際に入力された推論対象データに対する損失の情報を想定攻撃者が知ることができますか？

- ①想定攻撃者が判定結果しか得られない、あるいは判定結果すら得られないときは No です

②損失を入手できる関数を実行できるときなどは Yes です

質問 6－2. 推論処理の際に入力された推論対象データに対する勾配の情報を想定攻撃者が知ることができますか？

①想定攻撃者が判定結果しか得られない、あるいは判定結果すら得られないときは No です

②勾配を入手できる関数を実行できるときは Yes です

質問 6－3. 訓練済みモデルそのものを想定攻撃者が何らかの方法で入手することができますか？

7. 類似データセットの入手に関する質問

質問 7－1. AI システムの仕様として想定される入力データと類似のデータ（ほぼ同じ用途、ほぼ同じ種類・ジャンルのデータ）を、想定攻撃者が 1 個以上、何らかの手段で準備・入手できますか？

①類似データ 1 個とは AI が処理する最小単位のデータの集まりです。テーブルデータなら 1 行、画像なら 1 枚です。

質問 7－2. 【7-1】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 100 個程度以上、何らかの手段で準備・入手できますか？

質問 7－3. 【7-2】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 1,000 個程度以上、何らかの手段で準備・入手できますか？

質問 7－4. 【7-3】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 10,000 個程度以上、何らかの手段で準備・入手できますか？

8. 扱うデータに関する質問

質問 8－1. システムで扱っているデータはテーブルデータですか？

①実行する AI のタスクとしてテーブルデータをターゲットとしている場合は Yes、そうでなければ No を選択してください

②前処理があった場合には前処理後のデータがテーブルになっていたら当てはまります

II-7.4. 攻撃実施可能条件の満足状況の判定用テーブル

II-7.2.5 節で示した質問群への回答を元に、II-7.2 節で抽出した攻撃実施可能条件に合致しているかを判定するテーブルは表 II- 4 のようになる。なお、対応案が採用しにくい場合には、ツリーを成立させないための別の条件を算出してそちらを FALSE にすることを試みるべきである。また、それでも対応しにくい場合には、機械学習セキュリティ専用の対策を導入するべきであり、専門家への相談を要する。

表 II- 4. 攻撃実施可能条件満足状況判定用テーブル

| 攻撃実施可能条件 | 内容 | TRUE になる条件 | 判定結果 (TRUE/FALSE) | FALSE にする対策案 |
|----------|---------------------|--------------------------------|----------------------|---|
| 条件 1-1 | 訓練処理の自由な実行が可能 | 質問 1 – 1 A または質問 1 – 1 B が Yes | | 想定攻撃者が訓練処理を実行できないようにする |
| 条件 2-1 | 推論処理を 1 個以上で実行可能 | 質問 2 – 1 A または質問 2 – 1 B が Yes | | 想定攻撃者が推論処理を実行できないように設定する |
| 条件 2-2 | 推論処理を千個以上で実行可能 | 質問 2 – 2 A または質問 2 – 2 B が Yes | | 想定攻撃者がデータ 1,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-3 | 推論処理を 1 万個以上で実行可能 | 質問 2 – 3 A または質問 2 – 3 B が Yes | | 想定攻撃者がデータ 10,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-4 | 推論処理を 100 万個以上で実行可能 | 質問 2 – 4 A または質問 2 – 4 B が Yes | | 想定攻撃者がデータ 1,000,000 個以上に対して推論処理を実行できないように設定する |
| 条件 3-1 | 推論結果入手可能 | 質問 3 – 1 、または質問 3 – 2 が Yes | | 判定結果を想定攻撃者に提示しないようにする |
| 条件 3-2 | 確信度入手可能 | 質問 3 – 2 が Yes | | 判定結果の確信度を想定攻撃者に提示しないようにする |

| | | | | |
|--------|---------------------|-----------------|--|--|
| 条件 4-1 | データのメタ情報を入手可能 | 質問 4－1 が Yes | | 訓練関連データや推論対象データのメタ情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-2 | システムデータの統計情報が入手可能 | 質問 4－2 が Yes | | 訓練関連データや統計情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-3 | システムデータを 1 個以上入手可能 | 質問 4－3 が Yes | | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-4 | システムデータを千個以上入手可能 | 質問 4－4 が Yes | | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-5 | システムデータを 1 万個以上入手可能 | 質問 4－5 が Yes | | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |

| | | | | |
|--------|--------------------|--------------|--|---|
| 条件 4-6 | 訓練関連データを 1 個以上入手可能 | 質問 4－6 が Yes | | 訓練関連データを想定攻撃者が入手、推定できないようにする |
| 条件 4-7 | データの正解ラベルが入手可能 | 質問 4－7 が Yes | | システムの詳細な仕様を公開せず、想定攻撃者に真の正解ラベル (Ground Truth) を推測されないようにする |
| 条件 5-1 | 転移性を利用したモデルの開発 | 質問 5－1 が Yes | | 信頼できる入手先から得られるモデルのみをシステムに組み込む。 あるいは転移性を用いないようにする。 |
| 条件 6-1 | データに対する損失が入手可能 | 質問 6－1 が Yes | | 損失情報を想定攻撃者が入手できないようにする |
| 条件 6-2 | データに対する勾配が入手可能 | 質問 6－2 が Yes | | 勾配情報を想定攻撃者が入手できないようにする |
| 条件 6-3 | 訓練済みモデル自体を入手可能 | 質問 6－3 が Yes | | 訓練済みモデルを管理し、外部に流出しないようにする |
| 条件 7-1 | 類似データを 1 個以上入手可能 | 質問 7－1 が Yes | | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-2 | 類似データを 100 個以上入手可能 | 質問 7－2 が Yes | | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-3 | 類似データを 千個以上入手可能 | 質問 7－3 が Yes | | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |

| | | | | |
|--------|-------------------------|-----------------|--|--|
| 条件 7-4 | 類似データを 1万個以上入 手可能 | 質問 7－4 が Yes | | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 8-1 | 扱うデータが テーブルデー タ | 質問 8－1 が Yes | | テーブルデータを扱う AI システムなのか画像 データを扱う AI システ ムなのか、利用形態が適 切であるかを確認する |

II-7.5. AI リスク問診ツール

本章で説明した AI リスク問診の実現例をツールとして実装したものを本ガイドラインと同時に公開する。本ツールは Microsoft Excel で構成されている。I. AI の定義 と、II. 質問 のタブに書かれているシートを埋めると IV. 総合判定結果以降に判定結果が表示される。詳細はツールに付属の `readme`、及び、ツール内の説明を参照して頂きたい。

II-8. AI リスク問診の試行例

この章では II-7 章で紹介した実現例を利用した、機械学習システムの分析事例を紹介する。

II-8.1. 事例試行概要

[II-6]では、標識識別 AI に対する試行例が掲載されている。これに対して今回、策定委員において 3 つの事例について試行を実施した。事例を以下に示す。本章ではこれらの試行結果を紹介する。

- ・融資審査 AI
- ・プラント制御 AI
- ・性別・年齢推定 AI

II-8.1.1. 融資審査 AI

仕様：融資申込者が返済できるかどうかを予測する AI

金融情報と融資申込者の情報をデータ処理担当者が訓練してモデルを構築する。融資申込者の情報を金融担当者が入力し、AI が返済できるかどうかを予測（分類）する。推論結果は金融担当者のみが知ることができる。融資申込者には結果を見せない。

1. 想定攻撃者＝データ処理担当者

(i) 質問への回答

| 質問番号 | 質問 | 回答 |
|------|---|-----|
| 1-1A | 想定攻撃者の意思で訓練処理を行うことができる AI システムにおいて、自身の意図した数種類のデータで想定攻撃者が訓練処理を実行することができますか？ | Yes |
| 1-1B | 想定攻撃者でない人が訓練処理を行う AI システム、あるいは自動で訓練処理を行う AI システムにおいて、想定攻撃者が意図したデータを訓練データに混入できますか？ | - |
| 2-1A | 想定攻撃者が準備した推論対象データ 1 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-2A | 【2-1A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-3A | 【2-2A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000 個以上に対して推論処理を実行することができますか？ | Yes |

| | | |
|------|---|-----|
| | て、意図したデータ 10,000 個以上に対して推論処理を実行することができますか？ | |
| 2-4A | 【2-3A】が Yes であった場合、想定攻撃者は 【2-1A】において、意図したデータ 1,000,000 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-1B | 想定攻撃者が推論対象データを 1 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-2B | 【2-1B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 1,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-3B | 【2-2B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 10,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-4B | 【2-3B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 1,000,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 3-1 | 想定攻撃者にモデルの判定結果を提示しますか？ あるいは、想定攻撃者は判定結果を類推することができますか？ | Yes |
| 3-2 | モデルの確信度を一部でも想定攻撃者に提示しますか？ | Yes |
| 4-1 | 推論対象データや訓練データを作成できるフォーマット情報（データ形式、サイズの情報、画像の種類、データの次元等のメタ情報）を想定攻撃者は知ることができますか？ | Yes |
| 4-2 | AI システムに用いられた訓練データの統計情報を想定攻撃者は知ることができますか？ | Yes |
| 4-3 | AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ) を想定攻撃者が、1 個以上入手できますか？ | Yes |
| 4-4 | 【4-3】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or テストデータ or 推論対象データ) を想定攻撃者が、1,000 個程度以上入手できますか？ | Yes |
| 4-5 | 【4-4】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ) を想定攻撃者が、10,000 個程度以上入手できますか？ | Yes |
| 4-6 | 訓練関連データに関して、想定攻撃者が 1 個以上入手できま | Yes |

| | | |
|-----|---|-----|
| | すか？ | |
| 4-7 | AI システムへ入力したデータに対する正解ラベルを想定攻撃者が入手できますか？ | Yes |
| 5-1 | 外部や内部から入手した訓練済みモデルを一部でも流用し、転移性を利用して AI を構築していますか？ | No |
| 6-1 | 推論処理の際に入力された推論対象データに対する損失の情報を見たがる事ができますか？ | Yes |
| 6-2 | 推論処理の際に入力された推論対象データに対する勾配の情報を見たがる事ができますか？ | Yes |
| 6-3 | 訓練済みモデルそのものを想定攻撃者が何らかの方法で入手することができますか？ | Yes |
| 7-1 | AI システムの仕様として想定される入力データと類似のデータ（ほぼ同じ用途、ほぼ同じ種類・ジャンルのデータ）を、想定攻撃者が 1 個以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-2 | 【7-1】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 100 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-3 | 【7-2】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 1,000 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-4 | 【7-3】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 10,000 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 8-1 | システムで扱っているデータはテーブルデータですか？ | Yes |

(ii) 条件合致性判定

| 攻撃実施可能条件 | 内容 | TRUE になる条件 | 判定結果 (TRUE/FALSE) | FALSE にする対策案 |
|----------|------------------|--------------------------------|-------------------|--------------------------|
| 条件 1-1 | 訓練処理の自由な実行が可能 | 質問 1 - 1 A または質問 1 - 1 B が Yes | TRUE | 想定攻撃者が訓練処理を実行できないようにする |
| 条件 2-1 | 推論処理を 1 個以上で実行可能 | 質問 2 - 1 A または質問 2 - 1 B が Yes | TRUE | 想定攻撃者が推論処理を実行できないように設定する |

| | | | | |
|--------|---------------------|----------------------------|------|--|
| 条件 2-2 | 推論処理を千個以上で実行可能 | 質問 2－2 A または質問 2－2 B が Yes | TRUE | 想定攻撃者がデータ 1,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-3 | 推論処理を 1 万個以上で実行可能 | 質問 2－3 A または質問 2－3 B が Yes | TRUE | 想定攻撃者がデータ 10,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-4 | 推論処理を 100 万個以上で実行可能 | 質問 2－4 A または質問 2－4 B が Yes | TRUE | 想定攻撃者がデータ 1,000,000 個以上に対して推論処理を実行できないように設定する |
| 条件 3-1 | 推論結果入手可能 | 質問 3－1、または質問 3－2 が Yes | TRUE | 判定結果を想定攻撃者に提示しないようにする |
| 条件 3-2 | 確信度入手可能 | 質問 3－2 が Yes | TRUE | 判定結果の確信度を想定攻撃者に提示しないようにする |
| 条件 4-1 | データのメタ情報を入手可能 | 質問 4－1 が Yes | TRUE | 訓練関連データや推論対象データのメタ情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-2 | システムデータの統計情報入手可能 | 質問 4－2 が Yes | TRUE | 訓練関連データや統計情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-3 | システムデータを 1 個以上入手可能 | 質問 4－3 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |

| | | | | |
|--------|---------------------|--------------|-------|--|
| 条件 4-4 | システムデータを千個以上入手可能 | 質問 4－4 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-5 | システムデータを 1 万個以上入手可能 | 質問 4－5 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-6 | 訓練関連データを 1 個以上入手可能 | 質問 4－6 が Yes | TRUE | 訓練関連データを想定攻撃者が入手、推定できないようにする |
| 条件 4-7 | データの正解ラベルが入手可能 | 質問 4－7 が Yes | TRUE | システムの詳細な仕様を公開せず、想定攻撃者に真の正解ラベル (Ground Truth) を推測されないようにする |
| 条件 5-1 | 転移性を利用したモデルの開発 | 質問 5－1 が Yes | FALSE | 信頼できる入手先から得られるモデルのみをシステムに組み込む。 あるいは転移性を用いないようにする。 |
| 条件 6-1 | データに対する損失が入手可能 | 質問 6－1 が Yes | TRUE | 損失情報を想定攻撃者が入手できないようにする |

| | | | | |
|--------|--------------------|--------------|------|---|
| 条件 6-2 | データに対する勾配が入手可能 | 質問 6－2 が Yes | TRUE | 勾配情報を想定攻撃者が入手できないようにする |
| 条件 6-3 | 訓練済みモデル自体を入手可能 | 質問 6－3 が Yes | TRUE | 訓練済みモデルを管理し、外部に流出しないようにする |
| 条件 7-1 | 類似データを 1 個以上入手可能 | 質問 7－1 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-2 | 類似データを 100 個以上入手可能 | 質問 7－2 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-3 | 類似データを 千個以上入手可能 | 質問 7－3 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-4 | 類似データを 1 万個以上入手可能 | 質問 7－4 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 8-1 | 扱うデータが テーブルデータ | 質問 8－1 が Yes | TRUE | テーブルデータを扱う AI システムなのか画像データを扱う AI システムなのか、利用形態が適切であるかを確認する |

(iii) 成立した攻撃シナリオ（アタックツリーより判定）

回避攻撃（敵対的サンプル）：A1, A2, A3, A4

ポイズニング攻撃：P1, P3

モデル抽出攻撃：X1, X2, X3, X4, X6

モデルインバージョン攻撃：I1

メンバシップ推測攻撃：M1, M2, M3, M4, M5, M6, M7, M8

(iv) 分析結果

P2, X5 以外のすべてのシナリオで攻撃実施可能と判定された。データ処理担当者は AI 開発者と同等であり、権限が大きいためこのような結果となったと考えられる。

2. 想定攻撃者=金融担当者

(i) 質問への回答

| 質問番号 | 質問 | 回答 |
|------|---|-----|
| 1-1A | 想定攻撃者の意思で訓練処理を行うことができる AI システムにおいて、自身の意図した数種類のデータで想定攻撃者が訓練処理を実行することができますか？ | Yes |
| 1-1B | 想定攻撃者でない人が訓練処理を行う AI システム、あるいは自動で訓練処理を行う AI システムにおいて、想定攻撃者が意図したデータを訓練データに混入できますか？ | - |
| 2-1A | 想定攻撃者が準備した推論対象データ 1 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-2A | 【2-1A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-3A | 【2-2A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 10,000 個以上に対して推論処理を実行することができますか？ | No |
| 2-4A | 【2-3A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000,000 個以上に対して推論処理を実行することができますか？ | No |
| 2-1B | 想定攻撃者が推論対象データを 1 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-2B | 【2-1B】が Yes であった場合、想定攻撃者は【2-1B】において、推論対象データを 1,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-3B | 【2-2B】が Yes であった場合、想定攻撃者は【2-1B】において、推論対象データを 10,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-4B | 【2-3B】が Yes であった場合、想定攻撃者は【2-1B】において、推論対象データを 1,000,000 個以上置き換えたり、混入し | - |

| | | |
|-----|---|-----|
| | たりすることで、推論処理に通すことができますか？ | |
| 3-1 | 想定攻撃者にモデルの判定結果を提示しますか？ あるいは、想定攻撃者は判定結果を類推することができますか？ | Yes |
| 3-2 | モデルの確信度を一部でも想定攻撃者に提示しますか？ | Yes |
| 4-1 | 推論対象データや訓練データを作成できるフォーマット情報（データ形式、サイズの情報、画像の種類、データの次元等のメタ情報）を想定攻撃者は知ることができますか？ | Yes |
| 4-2 | AI システムに用いられた訓練データの統計情報を想定攻撃者は知ることができますか？ | Yes |
| 4-3 | AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ) を想定攻撃者が、1 個以上入手できますか？ | Yes |
| 4-4 | 【4-3】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or テストデータ or 推論対象データ) を想定攻撃者が、1,000 個程度以上入手できますか？ | Yes |
| 4-5 | 【4-4】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ) を想定攻撃者が、10,000 個程度以上入手できますか？ | Yes |
| 4-6 | 訓練関連データに関して、想定攻撃者が 1 個以上入手できますか？ | Yes |
| 4-7 | AI システムへ入力したデータに対する正解ラベルを想定攻撃者が入手できますか？ | Yes |
| 5-1 | 外部や内部から入手した訓練済みモデルを一部でも流用し、転移性を利用して AI を構築していますか？ | No |
| 6-1 | 推論処理の際に入力された推論対象データに対する損失の情報を想定攻撃者が知ることができますか？ | No |
| 6-2 | 推論処理の際に入力された推論対象データに対する勾配の情報を想定攻撃者が知ることができますか？ | No |
| 6-3 | 訓練済みモデルそのものを想定攻撃者が何らかの方法で入手することができますか？ | Yes |
| 7-1 | AI システムの仕様として想定される入力データと類似のデータ（ほぼ同じ用途、ほぼ同じ種類・ジャンルのデータ）を、想定攻撃者が 1 個以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-2 | 【7-1】が Yes であった場合、AI システムの仕様として想定す | Yes |

| | | |
|-----|---|-----|
| | るデータと類似のデータを、想定攻撃者が 100 個程度以上、何らかの手段で準備・入手できますか？ | |
| 7-3 | 【7-2】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 1,000 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-4 | 【7-3】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 10,000 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 8-1 | システムで扱っているデータはテーブルデータですか？ | Yes |

(ii) 条件合致性判定

| 攻撃実施可能条件 | 内容 | TRUE になる条件 | 判定結果 (TRUE/FALSE) | FALSE にする対策案 |
|----------|---------------------|--------------------------------|----------------------|---|
| 条件 1-1 | 訓練処理の自由な実行が可能 | 質問 1 - 1 A または質問 1 - 1 B が Yes | TRUE | 想定攻撃者が訓練処理を実行できないようにする |
| 条件 2-1 | 推論処理を 1 個以上で実行可能 | 質問 2 - 1 A または質問 2 - 1 B が Yes | TRUE | 想定攻撃者が推論処理を実行できないように設定する |
| 条件 2-2 | 推論処理を千個以上で実行可能 | 質問 2 - 2 A または質問 2 - 2 B が Yes | TRUE | 想定攻撃者がデータ 1,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-3 | 推論処理を 1 万個以上で実行可能 | 質問 2 - 3 A または質問 2 - 3 B が Yes | FALSE | 想定攻撃者がデータ 10,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-4 | 推論処理を 100 万個以上で実行可能 | 質問 2 - 4 A または質問 2 - 4 B が Yes | FALSE | 想定攻撃者がデータ 1,000,000 個以上に対して推論処理を実行できないように設定する |
| 条件 3-1 | 推論結果を入手可能 | 質問 3 - 1、または質問 3 - 2 が Yes | TRUE | 判定結果を想定攻撃者に提示しないようにする |

| | | | | |
|--------|---------------------|--------------|------|--|
| 条件 3-2 | 確信度を入手可能 | 質問 3－2 が Yes | TRUE | 判定結果の確信度を想定攻撃者に提示しないようにする |
| 条件 4-1 | データのメタ情報を入手可能 | 質問 4－1 が Yes | TRUE | 訓練関連データや推論対象データのメタ情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-2 | システムデータの統計情報が入手可能 | 質問 4－2 が Yes | TRUE | 訓練関連データや統計情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-3 | システムデータを 1 個以上入手可能 | 質問 4－3 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-4 | システムデータを千個以上入手可能 | 質問 4－4 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-5 | システムデータを 1 万個以上入手可能 | 質問 4－5 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入 |

| | | | | |
|--------|--------------------|--------------|-------|---|
| | | | | 手、推定できないようにする。 |
| 条件 4-6 | 訓練関連データを 1 個以上入手可能 | 質問 4－6 が Yes | TRUE | 訓練関連データを想定攻撃者が入手、推定できないようにする |
| 条件 4-7 | データの正解ラベルが入手可能 | 質問 4－7 が Yes | TRUE | システムの詳細な仕様を公開せず、想定攻撃者に真の正解ラベル (Ground Truth) を推測されないようにする |
| 条件 5-1 | 転移性を利用したモデルの開発 | 質問 5－1 が Yes | FALSE | 信頼できる入手先から得られるモデルのみをシステムに組み込む。 あるいは転移性を用いないようにする。 |
| 条件 6-1 | データに対する損失が入手可能 | 質問 6－1 が Yes | FALSE | 損失情報を想定攻撃者が入手できないようにする |
| 条件 6-2 | データに対する勾配が入手可能 | 質問 6－2 が Yes | FALSE | 勾配情報を想定攻撃者が入手できないようにする |
| 条件 6-3 | 訓練済みモデル自体を入手可能 | 質問 6－3 が Yes | TRUE | 訓練済みモデルを管理し、外部に流出しないようにする |
| 条件 7-1 | 類似データを 1 個以上入手可能 | 質問 7－1 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-2 | 類似データを 100 個以上入手可能 | 質問 7－2 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-3 | 類似データを 千個以上入手可能 | 質問 7－3 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |

| | | | | |
|--------|-------------------------|-------------------|------|--|
| 条件 7-4 | 類似データを 1万個以上入 手可能 | 質問 7 – 4 が Yes | TRUE | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 8-1 | 扱うデータが テーブルデー タ | 質問 8 – 1 が Yes | TRUE | テーブルデータを扱う AI システムなのか画像 データを扱う AI システ ムなのか、利用形態が適 切であるかを確認する |

(iii) 成立した攻撃シナリオ（アタックツリーより判定）

回避攻撃（敵対的サンプル）：A1, A2, A3, A4

ポイズニング攻撃：P1, P3

モデル抽出攻撃：X4, X6

モデルインバージョン攻撃：I1

メンバシップ推測攻撃：M3, M4, M5, M7, M8

(iv) 分析結果

金融担当者は自身の結果を金融情報に反映できるため質問 1 も Yes となった。ただし、データ処理担当者と異なり、業務外で大量の推論処理は実行できないため、このような結果となった。

3. 想定攻撃者＝第三者（融資申込者等）

(i) 質問への回答

| 質問番号 | 質問 | 回答 |
|------|---|-----|
| 1-1A | 想定攻撃者の意思で訓練処理を行うことができる AI システムにおいて、自身の意図した数種類のデータで想定攻撃者が訓練処理を実行することができますか？ | - |
| 1-1B | 想定攻撃者でない人が訓練処理を行う AI システム、あるいは自動で訓練処理を行う AI システムにおいて、想定攻撃者が意図したデータを訓練データに混入できますか？ | No |
| 2-1A | 想定攻撃者が準備した推論対象データ 1 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-2A | 【2-1A】が Yes であった場合、想定攻撃者は 【2-1A】において | No |

| | | |
|------|---|----|
| | て、意図したデータ 1,000 個以上に対して推論処理を実行することができますか？ | |
| 2-3A | 【2-2A】が Yes であった場合、想定攻撃者は 【2-1A】において、意図したデータ 10,000 個以上に対して推論処理を実行することができますか？ | No |
| 2-4A | 【2-3A】が Yes であった場合、想定攻撃者は 【2-1A】において、意図したデータ 1,000,000 個以上に対して推論処理を実行することができますか？ | No |
| 2-1B | 想定攻撃者が推論対象データを 1 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-2B | 【2-1B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 1,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-3B | 【2-2B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 10,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-4B | 【2-3B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 1,000,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 3-1 | 想定攻撃者にモデルの判定結果を提示しますか？ あるいは、想定攻撃者は判定結果を類推することができますか？ | No |
| 3-2 | モデルの確信度を一部でも想定攻撃者に提示しますか？ | No |
| 4-1 | 推論対象データや訓練データを作成できるフォーマット情報（データ形式、サイズの情報、画像の種類、データの次元等のメタ情報）を想定攻撃者は知ることができますか？ | No |
| 4-2 | AI システムに用いられた訓練データの統計情報を想定攻撃者は知ることができますか？ | No |
| 4-3 | AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ) を想定攻撃者が、1 個以上入手できますか？ | No |
| 4-4 | 【4-3】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or テストデータ or 推論対象データ) を想定攻撃者が、1,000 個程度以上入手できますか？ | No |
| 4-5 | 【4-4】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ | No |

| | | |
|-----|---|-----|
| | or 推論対象データ) を想定攻撃者が、10,000 個程度以上入手できますか？ | |
| 4-6 | 訓練関連データに関して、想定攻撃者が 1 個以上入手できますか？ | No |
| 4-7 | AI システムへ入力したデータに対する正解ラベルを想定攻撃者が入手できますか？ | No |
| 5-1 | 外部や内部から入手した訓練済みモデルを一部でも流用し、転移性を利用して AI を構築していますか？ | No |
| 6-1 | 推論処理の際に入力された推論対象データに対する損失の情報を想定攻撃者が知ることができますか？ | No |
| 6-2 | 推論処理の際に入力された推論対象データに対する勾配の情報を想定攻撃者が知ることができますか？ | No |
| 6-3 | 訓練済みモデルそのものを想定攻撃者が何らかの方法で入手することができますか？ | No |
| 7-1 | AI システムの仕様として想定される入力データと類似のデータ（ほぼ同じ用途、ほぼ同じ種類・ジャンルのデータ）を、想定攻撃者が 1 個以上、何らかの手段で準備・入手できますか？ | No |
| 7-2 | 【7-1】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 100 個程度以上、何らかの手段で準備・入手できますか？ | No |
| 7-3 | 【7-2】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 1,000 個程度以上、何らかの手段で準備・入手できますか？ | No |
| 7-4 | 【7-3】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 10,000 個程度以上、何らかの手段で準備・入手できますか？ | No |
| 8-1 | システムで扱っているデータはテーブルデータですか？ | Yes |

(ii) 条件合致性判定

| 攻撃実施可能条件 | 内容 | TRUE になる条件 | 判定結果 (TRUE/FALSE) | FALSE にする対策案 |
|----------|---------------|----------------------------|----------------------|------------------------|
| 条件 1-1 | 訓練処理の自由な実行が可能 | 質問 1-1 A または質問 1-1 B が Yes | FALSE | 想定攻撃者が訓練処理を実行できないようにする |

| | | | | |
|--------|---------------------|----------------------------|-------|--|
| 条件 2-1 | 推論処理を 1 個以上で実行可能 | 質問 2－1 A または質問 2－1 B が Yes | TRUE | 想定攻撃者が推論処理を実行できないように設定する |
| 条件 2-2 | 推論処理を千個以上で実行可能 | 質問 2－2 A または質問 2－2 B が Yes | FALSE | 想定攻撃者がデータ 1,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-3 | 推論処理を 1 万個以上で実行可能 | 質問 2－3 A または質問 2－3 B が Yes | FALSE | 想定攻撃者がデータ 10,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-4 | 推論処理を 100 万個以上で実行可能 | 質問 2－4 A または質問 2－4 B が Yes | FALSE | 想定攻撃者がデータ 1,000,000 個以上に対して推論処理を実行できないように設定する |
| 条件 3-1 | 推論結果入手可能 | 質問 3－1、または質問 3－2 が Yes | FALSE | 判定結果を想定攻撃者に提示しないようにする |
| 条件 3-2 | 確信度入手可能 | 質問 3－2 が Yes | FALSE | 判定結果の確信度を想定攻撃者に提示しないようにする |
| 条件 4-1 | データのメタ情報を入手可能 | 質問 4－1 が Yes | FALSE | 訓練関連データや推論対象データのメタ情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-2 | システムデータの統計情報を入手可能 | 質問 4－2 が Yes | FALSE | 訓練関連データや統計情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-3 | システムデータを 1 個以上入手可能 | 質問 4－3 が Yes | FALSE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入 |

| | | | | |
|--------|---------------------|--------------|-------|--|
| | | | | 手、推定できないようにする。 |
| 条件 4-4 | システムデータを千個以上入手可能 | 質問 4－4 が Yes | FALSE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-5 | システムデータを 1 万個以上入手可能 | 質問 4－5 が Yes | FALSE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-6 | 訓練関連データを 1 個以上入手可能 | 質問 4－6 が Yes | FALSE | 訓練関連データを想定攻撃者が入手、推定できないようにする |
| 条件 4-7 | データの正解ラベルが入手可能 | 質問 4－7 が Yes | FALSE | システムの詳細な仕様を公開せず、想定攻撃者に真の正解ラベル (Ground Truth) を推測されないようにする |
| 条件 5-1 | 転移性を利用したモデルの開発 | 質問 5－1 が Yes | FALSE | 信頼できる入手先から得られるモデルのみをシステムに組み込む。 あるいは転移性を用いないようにする。 |
| 条件 6-1 | データに対する損失が入手可能 | 質問 6－1 が Yes | FALSE | 損失情報を想定攻撃者が入手できないようにする |

| | | | | |
|--------|--------------------|--------------|-------|---|
| 条件 6-2 | データに対する勾配が入手可能 | 質問 6－2 が Yes | FALSE | 勾配情報を想定攻撃者が入手できないようにする |
| 条件 6-3 | 訓練済みモデル自体を入手可能 | 質問 6－3 が Yes | FALSE | 訓練済みモデルを管理し、外部に流出しないようにする |
| 条件 7-1 | 類似データを 1 個以上入手可能 | 質問 7－1 が Yes | FALSE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-2 | 類似データを 100 個以上入手可能 | 質問 7－2 が Yes | FALSE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-3 | 類似データを 千個以上入手可能 | 質問 7－3 が Yes | FALSE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-4 | 類似データを 1 万個以上入手可能 | 質問 7－4 が Yes | FALSE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 8-1 | 扱うデータが テーブルデータ | 質問 8－1 が Yes | TRUE | テーブルデータを扱う AI システムなのか画像データを扱う AI システムなのか、利用形態が適切であるかを確認する |

(iii) 成立した攻撃シナリオ（アタックツリーより判定）

回避攻撃（敵対的サンプル）：なし

ポイズニング攻撃：なし

モデル抽出攻撃：なし

モデルインバージョン攻撃：なし

メンバシップ推測攻撃：なし

(iv) 分析結果

融資申込者を含む第三者は、自身が申し込むことで AI を間接的に実行させることはでき、自身のデータを入力することができる。しかし結果を入手することはできない。このため、想定した全ての攻撃シナリオについて実施困難と判定された。

II-8.1.2. プラント制御 AI

仕様：プラントの酸素供給量の判断を行う AI

センサから得た情報を元に酸素供給量を判断する。訓練はプラント関係者が行う。推論は定期的に行われ、人間は関与しない。今回は第三者により攻撃が実施可能かどうかを分析した。第三者はセンサを自分で作成したものに置き換えて AI システムにデータを送る想定とした。また、プラントは 1 日 1 回見回りに来ることを想定し、最大でも 24 時間しか異常なデータは流せないとした。

1. 想定攻撃者＝第三者

(i) 質問への回答

| 質問番号 | 質問 | 回答 |
|------|---|-----|
| 1-1A | 想定攻撃者の意思で訓練処理を行うことができる AI システムにおいて、自身の意図した数種類のデータで想定攻撃者が訓練処理を実行することができますか？ | - |
| 1-1B | 想定攻撃者でない人が訓練処理を行う AI システム、あるいは自動で訓練処理を行う AI システムにおいて、想定攻撃者が意図したデータを訓練データに混入できますか？ | No |
| 2-1A | 想定攻撃者が準備した推論対象データ 1 個以上に対して推論処理を実行することができますか？ | - |
| 2-2A | 【2-1A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000 個以上に対して推論処理を実行することができますか？ | - |
| 2-3A | 【2-2A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 10,000 個以上に対して推論処理を実行することができますか？ | - |
| 2-4A | 【2-3A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000,000 個以上に対して推論処理を実行することができますか？ | - |
| 2-1B | 想定攻撃者が推論対象データを 1 個以上置き換えたり、混入 | Yes |

| | | |
|------|---|-----|
| | したりすることで、推論処理に通すことができますか？ | |
| 2-2B | 【2-1B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 1,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | Yes |
| 2-3B | 【2-2B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 10,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | Yes |
| 2-4B | 【2-3B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 1,000,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | No |
| 3-1 | 想定攻撃者にモデルの判定結果を提示しますか？ あるいは、想定攻撃者は判定結果を類推することができますか？ | No |
| 3-2 | モデルの確信度を一部でも想定攻撃者に提示しますか？ | No |
| 4-1 | 推論対象データや訓練データを作成できるフォーマット情報（データ形式、サイズの情報、画像の種類、データの次元等のメタ情報）を想定攻撃者は知ることができますか？ | No |
| 4-2 | AI システムに用いられた訓練データの統計情報を想定攻撃者は知ることができますか？ | No |
| 4-3 | AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ) を想定攻撃者が、1 個以上入手できますか？ | Yes |
| 4-4 | 【4-3】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or テストデータ or 推論対象データ) を想定攻撃者が、1,000 個程度以上入手できますか？ | Yes |
| 4-5 | 【4-4】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ) を想定攻撃者が、10,000 個程度以上入手できますか？ | Yes |
| 4-6 | 訓練関連データに関して、想定攻撃者が 1 個以上入手できますか？ | No |
| 4-7 | AI システムへ入力したデータに対する正解ラベルを想定攻撃者が入手できますか？ | No |
| 5-1 | 外部や内部から入手した訓練済みモデルを一部でも流用し、転移性を利用して AI を構築していますか？ | No |
| 6-1 | 推論処理の際に入力された推論対象データに対する損失の情 | No |

| | | |
|-----|---|-----|
| | 報を想定攻撃者が知ることができますか？ | |
| 6-2 | 推論処理の際に入力された推論対象データに対する勾配の情報 報を想定攻撃者が知ることができますか？ | No |
| 6-3 | 訓練済みモデルそのものを想定攻撃者が何らかの方法で入手 することができますか？ | No |
| 7-1 | AI システムの仕様として想定される入力データと類似のデータ（ほぼ同じ用途、ほぼ同じ種類・ジャンルのデータ）を、想定攻撃者が 1 個以上、何らかの手段で準備・入手できますか？ | No |
| 7-2 | 【7-1】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 100 個程度以上、何らかの手段で準備・入手できますか？ | No |
| 7-3 | 【7-2】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 1,000 個程度以上、何らかの手段で準備・入手できますか？ | No |
| 7-4 | 【7-3】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 10,000 個程度以上、何らかの手段で準備・入手できますか？ | No |
| 8-1 | システムで扱っているデータはテーブルデータですか？ | Yes |

(ii) 条件合致性判定

| 攻撃実施可能条件 | 内容 | TRUE になる条件 | 判定結果 (TRUE/FALSE) | FALSE にする対策案 |
|----------|-------------------|-------------------------------|-------------------|--|
| 条件 1-1 | 訓練処理の自由な実行が可能 | 質問 1－1 A または質問 1－1 B が Yes | FALSE | 想定攻撃者が訓練処理を実行できないようにする |
| 条件 2-1 | 推論処理を 1 個以上で実行可能 | 質問 2－1 A または質問 2－1 B が Yes | TRUE | 想定攻撃者が推論処理を実行できないように設定する |
| 条件 2-2 | 推論処理を千個以上で実行可能 | 質問 2－2 A または質問 2－2 B が Yes | TRUE | 想定攻撃者がデータ 1,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-3 | 推論処理を 1 万個以上で実行可能 | 質問 2－3 A または質問 2－3 B が Yes | TRUE | 想定攻撃者がデータ 10,000 個以上に対して推論処理を実行できないように設定する |

| | | | | |
|--------|---------------------|--------------------------------|-------|--|
| 条件 2-4 | 推論処理を 100 万個以上で実行可能 | 質問 2 – 4 A または質問 2 – 4 B が Yes | FALSE | 想定攻撃者がデータ 1,000,000 個以上に対して推論処理を実行できないように設定する |
| 条件 3-1 | 推論結果を入手可能 | 質問 3 – 1、または質問 3 – 2 が Yes | FALSE | 判定結果を想定攻撃者に提示しないようにする |
| 条件 3-2 | 確信度を入手可能 | 質問 3 – 2 が Yes | FALSE | 判定結果の確信度を想定攻撃者に提示しないようにする |
| 条件 4-1 | データのメタ情報を入手可能 | 質問 4 – 1 が Yes | FALSE | 訓練関連データや推論対象データのメタ情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-2 | システムデータの統計情報を入手可能 | 質問 4 – 2 が Yes | FALSE | 訓練関連データや統計情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-3 | システムデータを 1 個以上入手可能 | 質問 4 – 3 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-4 | システムデータを千個以上入手可能 | 質問 4 – 4 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |

| | | | | |
|--------|---------------------|--------------|-------|--|
| 条件 4-5 | システムデータを 1 万個以上入手可能 | 質問 4－5 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-6 | 訓練関連データを 1 個以上入手可能 | 質問 4－6 が Yes | FALSE | 訓練関連データを想定攻撃者が入手、推定できないようにする |
| 条件 4-7 | データの正解ラベルが入手可能 | 質問 4－7 が Yes | FALSE | システムの詳細な仕様を公開せず、想定攻撃者に真の正解ラベル (Ground Truth) を推測されないようにする |
| 条件 5-1 | 転移性を利用したモデルの開発 | 質問 5－1 が Yes | FALSE | 信頼できる入手先から得られるモデルのみをシステムに組み込む。 あるいは転移性を用いないようにする。 |
| 条件 6-1 | データに対する損失が入手可能 | 質問 6－1 が Yes | FALSE | 損失情報を想定攻撃者が入手できないようにする |
| 条件 6-2 | データに対する勾配が入手可能 | 質問 6－2 が Yes | FALSE | 勾配情報を想定攻撃者が入手できないようにする |
| 条件 6-3 | 訓練済みモデル自身を入手可能 | 質問 6－3 が Yes | FALSE | 訓練済みモデルを管理し、外部に流出しないようにする |
| 条件 7-1 | 類似データを 1 個以上入手可能 | 質問 7－1 が Yes | FALSE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |

| | | | | |
|--------|---------------------------|-----------------|-------|--|
| 条件 7-2 | 類似データを 100 個以上入 手可能 | 質問 7－2 が Yes | FALSE | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 7-3 | 類似データを 千個以上入手 可能 | 質問 7－3 が Yes | FALSE | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 7-4 | 類似データを 1 万個以上入 手可能 | 質問 7－4 が Yes | FALSE | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 8-1 | 扱うデータが テーブルデー タ | 質問 8－1 が Yes | TRUE | テーブルデータを扱う AI システムなのか画像 データを扱う AI シス テムなのか、利用形態が適 切であるかを確認する |

(iii) 成立した攻撃シナリオ（アタックツリーより判定）

回避攻撃（敵対的サンプル）：なし

ポイズニング攻撃：なし

モデル抽出攻撃：なし

モデルインバージョン攻撃：なし

メンバシップ推測攻撃：なし

(iv) 分析結果

第三者はプラントの外部センサを細工し、好きなデータを入力できると想定。ただし、結果を入手できないため、想定した全ての攻撃シナリオが攻撃困難と判定された。

II-8.1.3. 性別・年齢推定 AI

仕様：撮影した映像に含まれる人物の性別・年齢を予測する AI

物体認識と性別・年齢判定の 2 つの AI の組み合わせとなる。訓練はカメラ映像に対し
て人手でラベル付けして行う。訓練は信頼できる人が行う。推論は店舗内を録画した
カメラ映像から画像を抜き取り、モデルに入力することで行う。結果は分析者のみ

が知ることができ、販売促進活動などに利用する。

1. 想定攻撃者=AI システム開発者

(i) 質問への回答

| 質問番号 | 質問 | 回答 |
|------|--|-----|
| 1-1A | 想定攻撃者の意思で訓練処理を行うことができる AI システムにおいて、自身の意図した数種類のデータで想定攻撃者が訓練処理を実行することができますか？ | Yes |
| 1-1B | 想定攻撃者でない人が訓練処理を行う AI システム、あるいは自動で訓練処理を行う AI システムにおいて、想定攻撃者が意図したデータを訓練データに混入できますか？ | - |
| 2-1A | 想定攻撃者が準備した推論対象データ 1 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-2A | 【2-1A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-3A | 【2-2A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 10,000 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-4A | 【2-3A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000,000 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-1B | 想定攻撃者が推論対象データを 1 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-2B | 【2-1B】が Yes であった場合、想定攻撃者は【2-1B】において、推論対象データを 1,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-3B | 【2-2B】が Yes であった場合、想定攻撃者は【2-1B】において、推論対象データを 10,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-4B | 【2-3B】が Yes であった場合、想定攻撃者は【2-1B】において、推論対象データを 1,000,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 3-1 | 想定攻撃者にモデルの判定結果を提示しますか？ あるいは、想定攻撃者は判定結果を類推することができますか？ | Yes |

| | | |
|-----|---|-----|
| 3-2 | モデルの確信度を一部でも想定攻撃者に提示しますか？ | Yes |
| 4-1 | 推論対象データや訓練データを作成できるフォーマット情報（データ形式、サイズの情報、画像の種類、データの次元等のメタ情報）を想定攻撃者は知ることができますか？ | Yes |
| 4-2 | AI システムに用いられた訓練データの統計情報を想定攻撃者は知ることができますか？ | Yes |
| 4-3 | AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ) を想定攻撃者が、1 個以上入手できますか？ | Yes |
| 4-4 | 【4-3】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or テストデータ or 推論対象データ) を想定攻撃者が、1,000 個程度以上入手できますか？ | Yes |
| 4-5 | 【4-4】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ) を想定攻撃者が、10,000 個程度以上入手できますか？ | Yes |
| 4-6 | 訓練関連データに関して、想定攻撃者が 1 個以上入手できますか？ | Yes |
| 4-7 | AI システムへ入力したデータに対する正解ラベルを想定攻撃者が入手できますか？ | Yes |
| 5-1 | 外部や内部から入手した訓練済みモデルを一部でも流用し、転移性を利用して AI を構築していますか？ | Yes |
| 6-1 | 推論処理の際に入力された推論対象データに対する損失の情報を想定攻撃者が知ることができますか？ | Yes |
| 6-2 | 推論処理の際に入力された推論対象データに対する勾配の情報を想定攻撃者が知ることができますか？ | Yes |
| 6-3 | 訓練済みモデルそのものを想定攻撃者が何らかの方法で入手することができますか？ | Yes |
| 7-1 | AI システムの仕様として想定される入力データと類似のデータ（ほぼ同じ用途、ほぼ同じ種類・ジャンルのデータ）を、想定攻撃者が 1 個以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-2 | 【7-1】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 100 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-3 | 【7-2】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 1,000 個程度以上、 | Yes |

| | | |
|-----|---|-----|
| | 何らかの手段で準備・入手できますか？ | |
| 7-4 | 【7-3】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 10,000 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 8-1 | システムで扱っているデータはテーブルデータですか？ | No |

(ii) 条件合致性判定

| 攻撃実施可能条件 | 内容 | TRUE になる条件 | 判定結果 (TRUE/FALSE) | FALSE にする対策案 |
|----------|---------------------|--------------------------------|----------------------|---|
| 条件 1-1 | 訓練処理の自由な実行が可能 | 質問 1 - 1 A または質問 1 - 1 B が Yes | TRUE | 想定攻撃者が訓練処理を実行できないようにする |
| 条件 2-1 | 推論処理を 1 個以上で実行可能 | 質問 2 - 1 A または質問 2 - 1 B が Yes | TRUE | 想定攻撃者が推論処理を実行できないように設定する |
| 条件 2-2 | 推論処理を千個以上で実行可能 | 質問 2 - 2 A または質問 2 - 2 B が Yes | TRUE | 想定攻撃者がデータ 1,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-3 | 推論処理を 1 万個以上で実行可能 | 質問 2 - 3 A または質問 2 - 3 B が Yes | TRUE | 想定攻撃者がデータ 10,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-4 | 推論処理を 100 万個以上で実行可能 | 質問 2 - 4 A または質問 2 - 4 B が Yes | TRUE | 想定攻撃者がデータ 1,000,000 個以上に対して推論処理を実行できないように設定する |
| 条件 3-1 | 推論結果を入手可能 | 質問 3 - 1、または質問 3 - 2 が Yes | TRUE | 判定結果を想定攻撃者に提示しないようにする |
| 条件 3-2 | 確信度を入手可能 | 質問 3 - 2 が Yes | TRUE | 判定結果の確信度を想定攻撃者に提示しないようにする |
| 条件 4-1 | データのメタ情報を入手可能 | 質問 4 - 1 が Yes | TRUE | 訓練関連データや推論対象データのメタ情報を想 |

| | | | | |
|--------|---------------------|--------------|------|--|
| | | | | 定攻撃者が入手、推定できないようにする |
| 条件 4-2 | システムデータの統計情報が入手可能 | 質問 4－2 が Yes | TRUE | 訓練関連データや統計情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-3 | システムデータを 1 個以上入手可能 | 質問 4－3 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-4 | システムデータを千個以上入手可能 | 質問 4－4 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-5 | システムデータを 1 万個以上入手可能 | 質問 4－5 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-6 | 訓練関連データを 1 個以上入手可能 | 質問 4－6 が Yes | TRUE | 訓練関連データを想定攻撃者が入手、推定できないようにする |

| | | | | |
|--------|--------------------|--------------|------|--|
| 条件 4-7 | データの正解ラベルが入手可能 | 質問 4－7 が Yes | TRUE | システムの詳細な仕様を公開せず、想定攻撃者に真の正解ラベル(Ground Truth) を推測されないようにする |
| 条件 5-1 | 転移性を利用したモデルの開発 | 質問 5－1 が Yes | TRUE | 信頼できる入手先から得られるモデルのみをシステムに組み込む。 あるいは転移性を用いないようにする。 |
| 条件 6-1 | データに対する損失が入手可能 | 質問 6－1 が Yes | TRUE | 損失情報を想定攻撃者が入手できないようにする |
| 条件 6-2 | データに対する勾配が入手可能 | 質問 6－2 が Yes | TRUE | 勾配情報を想定攻撃者が入手できないようにする |
| 条件 6-3 | 訓練済みモデル自身を入手可能 | 質問 6－3 が Yes | TRUE | 訓練済みモデルを管理し、外部に流出しないようにする |
| 条件 7-1 | 類似データを 1 個以上入手可能 | 質問 7－1 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-2 | 類似データを 100 個以上入手可能 | 質問 7－2 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-3 | 類似データを 千個以上入手可能 | 質問 7－3 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-4 | 類似データを 1 万個以上入手可能 | 質問 7－4 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |

| | | | | |
|--------|-------------------|-------------------|-------|--|
| 条件 8-1 | 扱うデータが テーブルデータ | 質問 8 – 1 が Yes | FALSE | テーブルデータを扱う AI システムなのか画像 データを扱う AI シス テムなのか、利用形態が適 切であるかを確認する |
|--------|-------------------|-------------------|-------|--|

(iii) 成立した攻撃シナリオ（アタックツリーより判定）

回避攻撃（敵対的サンプル）：A1, A2, A3, A4

ポイズニング攻撃：P1, P2, P3

モデル抽出攻撃：X1, X2, X3, X5, X6

モデルインバージョン攻撃：I1

メンバシップ推測攻撃：M1, M2, M3, M4, M5, M6, M7, M8

(iv) 分析結果

X4 以外のすべてのシナリオで攻撃実施可能と判定された。開発者は権限が大きいためこのような結果となったと考えられる。X4 は AI が扱うデータがテーブルデータであった際のシナリオのため、こちらは不成立となった。

2. 想定攻撃者＝ラベル付け担当者

(i) 質問への回答

| 質問番号 | 質問 | 回答 |
|------|---|-----|
| 1-1A | 想定攻撃者の意思で訓練処理を行うことができる AI システムにおいて、自身の意図した数種類のデータで想定攻撃者が訓練処理を実行することができますか？ | - |
| 1-1B | 想定攻撃者でない人が訓練処理を行う AI システム、あるいは自動で訓練処理を行う AI システムにおいて、想定攻撃者が意図したデータを訓練データに混入できますか？ | No |
| 2-1A | 想定攻撃者が準備した推論対象データ 1 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-2A | 【2-1A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-3A | 【2-2A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 10,000 個以上に対して推論処理を実行す | Yes |

| | | |
|------|--|-----|
| | ることができますか？ | |
| 2-4A | 【2-3A】が Yes であった場合、想定攻撃者は 【2-1A】において、意図したデータ 1,000,000 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-1B | 想定攻撃者が推論対象データを 1 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-2B | 【2-1B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 1,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-3B | 【2-2B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 10,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-4B | 【2-3B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 1,000,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 3-1 | 想定攻撃者にモデルの判定結果を提示しますか？ あるいは、想定攻撃者は判定結果を類推することができますか？ | No |
| 3-2 | モデルの確信度を一部でも想定攻撃者に提示しますか？ | No |
| 4-1 | 推論対象データや訓練データを作成できるフォーマット情報（データ形式、サイズの情報、画像の種類、データの次元等のメタ情報）を想定攻撃者は知ることができますか？ | No |
| 4-2 | AI システムに用いられた訓練データの統計情報を想定攻撃者は知ることができますか？ | No |
| 4-3 | AI システムへの入力データそのもの（訓練データ or バリデーションデータ or テストデータ or 推論対象データ）を想定攻撃者が、1 個以上入手できますか？ | No |
| 4-4 | 【4-3】が Yes であった場合、AI システムへの入力データそのもの（訓練データ or テストデータ or 推論対象データ）を想定攻撃者が、1,000 個程度以上入手できますか？ | No |
| 4-5 | 【4-4】が Yes であった場合、AI システムへの入力データそのもの（訓練データ or バリデーションデータ or テストデータ or 推論対象データ）を想定攻撃者が、10,000 個程度以上入手できますか？ | No |
| 4-6 | 訓練関連データに関して、想定攻撃者が 1 個以上入手できますか？ | No |

| | | |
|-----|---|-----|
| 4-7 | AI システムへ入力したデータに対する正解ラベルを想定攻撲者が入手できますか？ | Yes |
| 5-1 | 外部や内部から入手した訓練済みモデルを一部でも流用し、転移性を利用して AI を構築していますか？ | Yes |
| 6-1 | 推論処理の際に入力された推論対象データに対する損失の情報を想定攻撃者が知ることができますか？ | No |
| 6-2 | 推論処理の際に入力された推論対象データに対する勾配の情報を想定攻撃者が知ることができますか？ | No |
| 6-3 | 訓練済みモデルそのものを想定攻撃者が何らかの方法で入手することができますか？ | No |
| 7-1 | AI システムの仕様として想定される入力データと類似のデータ（ほぼ同じ用途、ほぼ同じ種類・ジャンルのデータ）を、想定攻撃者が 1 個以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-2 | 【7-1】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 100 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-3 | 【7-2】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 1,000 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-4 | 【7-3】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 10,000 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 8-1 | システムで扱っているデータはテーブルデータですか？ | No |

(ii) 条件合致性判定

| 攻撃実施可能条件 | 内容 | TRUE になる条件 | 判定結果 (TRUE/FALSE) | FALSE にする対策案 |
|----------|------------------|--------------------------------|-------------------|--------------------------|
| 条件 1-1 | 訓練処理の自由な実行が可能 | 質問 1 - 1 A または質問 1 - 1 B が Yes | FALSE | 想定攻撃者が訓練処理を実行できないようにする |
| 条件 2-1 | 推論処理を 1 個以上で実行可能 | 質問 2 - 1 A または質問 2 - 1 B が Yes | TRUE | 想定攻撃者が推論処理を実行できないように設定する |
| 条件 2-2 | 推論処理を千個以上で実行可能 | 質問 2 - 2 A または質問 2 - 2 B が Yes | TRUE | 想定攻撃者がデータ 1,000 個以上に対して推 |

| | | | | |
|--------|---------------------|--------------------------------|-------|--|
| | | | | 論処理を実行できないよう設定する |
| 条件 2-3 | 推論処理を 1 万個以上で実行可能 | 質問 2 – 3 A または質問 2 – 3 B が Yes | TRUE | 想定攻撃者がデータ 10,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-4 | 推論処理を 100 万個以上で実行可能 | 質問 2 – 4 A または質問 2 – 4 B が Yes | TRUE | 想定攻撃者がデータ 1,000,000 個以上に対して推論処理を実行できないように設定する |
| 条件 3-1 | 推論結果を入手可能 | 質問 3 – 1、または質問 3 – 2 が Yes | FALSE | 判定結果を想定攻撃者に提示しないようにする |
| 条件 3-2 | 確信度を入手可能 | 質問 3 – 2 が Yes | FALSE | 判定結果の確信度を想定攻撃者に提示しないようにする |
| 条件 4-1 | データのメタ情報を入手可能 | 質問 4 – 1 が Yes | FALSE | 訓練関連データや推論対象データのメタ情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-2 | システムデータの統計情報を入手可能 | 質問 4 – 2 が Yes | FALSE | 訓練関連データや統計情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-3 | システムデータを 1 個以上入手可能 | 質問 4 – 3 が Yes | FALSE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-4 | システムデータを千個以上入手可能 | 質問 4 – 4 が Yes | FALSE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 |

| | | | | |
|--------|---------------------|--------------|-------|--|
| | | | | また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-5 | システムデータを 1 万個以上入手可能 | 質問 4－5 が Yes | FALSE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-6 | 訓練関連データを 1 個以上入手可能 | 質問 4－6 が Yes | FALSE | 訓練関連データを想定攻撃者が入手、推定できないようにする |
| 条件 4-7 | データの正解ラベルが入手可能 | 質問 4－7 が Yes | TRUE | システムの詳細な仕様を公開せず、想定攻撃者に真の正解ラベル (Ground Truth) を推測されないようにする |
| 条件 5-1 | 転移性を利用したモデルの開発 | 質問 5－1 が Yes | TRUE | 信頼できる入手先から得られるモデルのみをシステムに組み込む。 あるいは転移性を用いないようにする。 |
| 条件 6-1 | データに対する損失が入手可能 | 質問 6－1 が Yes | FALSE | 損失情報を想定攻撃者が入手できないようにする |
| 条件 6-2 | データに対する勾配が入手可能 | 質問 6－2 が Yes | FALSE | 勾配情報を想定攻撃者が入手できないようにする |
| 条件 6-3 | 訓練済みモデル自体を入手可能 | 質問 6－3 が Yes | FALSE | 訓練済みモデルを管理し、外部に流出しないようにする |

| | | | | |
|--------|---------------------------|-----------------|-------|--|
| 条件 7-1 | 類似データを 1 個以上入手 可能 | 質問 7－1 が Yes | TRUE | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 7-2 | 類似データを 100 個以上入 手可能 | 質問 7－2 が Yes | TRUE | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 7-3 | 類似データを 千個以上入手 可能 | 質問 7－3 が Yes | TRUE | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 7-4 | 類似データを 1 万個以上入 手可能 | 質問 7－4 が Yes | TRUE | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 8-1 | 扱うデータが テーブルデー タ | 質問 8－1 が Yes | FALSE | テーブルデータを扱う AI システムなのか画像 データを扱う AI システ ムなのか、利用形態が適 切であるかを確認する |

(iii) 成立した攻撃シナリオ（アタックツリーより判定）

回避攻撃（敵対的サンプル）：なし

ポイズニング攻撃：P2

モデル抽出攻撃：なし

モデルインバージョン攻撃：なし

メンバシップ推測攻撃：なし

(iv) 分析結果

P2 のみ成立となった。想定攻撃者は訓練データにラベルを付けるだけであり、出力結果を得ることができない。よって大半の攻撃は実施困難となった。P2 はモデルの構造に起因しており、モデルを流用して AI を構築している際に、流用部分がはじめから汚染されている可能性を懸念してこの判断となっている。

3. 想定攻撃者＝分析者

(i) 質問への回答

| 質問番号 | 質問 | 回答 |
|------|--|-----|
| 1-1A | 想定攻撃者の意思で訓練処理を行うことができる AI システムにおいて、自身の意図した数種類のデータで想定攻撃者が訓練処理を実行することができますか？ | - |
| 1-1B | 想定攻撃者でない人が訓練処理を行う AI システム、あるいは自動で訓練処理を行う AI システムにおいて、想定攻撃者が意図したデータを訓練データに混入できますか？ | No |
| 2-1A | 想定攻撃者が準備した推論対象データ 1 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-2A | 【2-1A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-3A | 【2-2A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 10,000 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-4A | 【2-3A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000,000 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-1B | 想定攻撃者が推論対象データを 1 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-2B | 【2-1B】が Yes であった場合、想定攻撃者は【2-1B】において、推論対象データを 1,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-3B | 【2-2B】が Yes であった場合、想定攻撃者は【2-1B】において、推論対象データを 10,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-4B | 【2-3B】が Yes であった場合、想定攻撃者は【2-1B】において、推論対象データを 1,000,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 3-1 | 想定攻撃者にモデルの判定結果を提示しますか？ あるいは、想定攻撃者は判定結果を類推することができますか？ | Yes |
| 3-2 | モデルの確信度を一部でも想定攻撃者に提示しますか？ | Yes |

| | | |
|-----|---|-----|
| 4-1 | 推論対象データや訓練データを作成できるフォーマット情報（データ形式、サイズの情報、画像の種類、データの次元等のメタ情報）を想定攻撃者は知ることができますか？ | No |
| 4-2 | AI システムに用いられた訓練データの統計情報を想定攻撃者は知ることができますか？ | No |
| 4-3 | AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ) を想定攻撃者が、1 個以上入手できますか？ | Yes |
| 4-4 | 【4-3】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or テストデータ or 推論対象データ) を想定攻撃者が、1,000 個程度以上入手できますか？ | No |
| 4-5 | 【4-4】が Yes であった場合、AI システムへの入力データそのもの(訓練データ or バリデーションデータ or テストデータ or 推論対象データ) を想定攻撃者が、10,000 個程度以上入手できますか？ | No |
| 4-6 | 訓練関連データに関して、想定攻撃者が 1 個以上入手できますか？ | Yes |
| 4-7 | AI システムへ入力したデータに対する正解ラベルを想定攻撃者が入手できますか？ | No |
| 5-1 | 外部や内部から入手した訓練済みモデルを一部でも流用し、転移性を利用して AI を構築していますか？ | Yes |
| 6-1 | 推論処理の際に入力された推論対象データに対する損失の情報を想定攻撃者が知ることができますか？ | No |
| 6-2 | 推論処理の際に入力された推論対象データに対する勾配の情報を想定攻撃者が知ることができますか？ | No |
| 6-3 | 訓練済みモデルそのものを想定攻撃者が何らかの方法で入手することができますか？ | No |
| 7-1 | AI システムの仕様として想定される入力データと類似のデータ（ほぼ同じ用途、ほぼ同じ種類・ジャンルのデータ）を、想定攻撃者が 1 個以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-2 | 【7-1】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 100 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-3 | 【7-2】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 1,000 個程度以上、何らかの手段で準備・入手できますか？ | Yes |

| | | |
|-----|---|-----|
| 7-4 | 【7-3】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 10,000 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 8-1 | システムで扱っているデータはテーブルデータですか？ | No |

(ii) 条件合致性判定

| 攻撃実施可能条件 | 内容 | TRUE になる条件 | 判定結果 (TRUE/FALSE) | FALSE にする対策案 |
|----------|---------------------|--------------------------------|----------------------|---|
| 条件 1-1 | 訓練処理の自由な実行が可能 | 質問 1 – 1 A または質問 1 – 1 B が Yes | FALSE | 想定攻撃者が訓練処理を実行できないようにする |
| 条件 2-1 | 推論処理を 1 個以上で実行可能 | 質問 2 – 1 A または質問 2 – 1 B が Yes | TRUE | 想定攻撃者が推論処理を実行できないように設定する |
| 条件 2-2 | 推論処理を千個以上で実行可能 | 質問 2 – 2 A または質問 2 – 2 B が Yes | TRUE | 想定攻撃者がデータ 1,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-3 | 推論処理を 1 万個以上で実行可能 | 質問 2 – 3 A または質問 2 – 3 B が Yes | TRUE | 想定攻撃者がデータ 10,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-4 | 推論処理を 100 万個以上で実行可能 | 質問 2 – 4 A または質問 2 – 4 B が Yes | TRUE | 想定攻撃者がデータ 1,000,000 個以上に対して推論処理を実行できないように設定する |
| 条件 3-1 | 推論結果を入手可能 | 質問 3 – 1、または質問 3 – 2 が Yes | TRUE | 判定結果を想定攻撃者に提示しないようにする |
| 条件 3-2 | 確信度を入手可能 | 質問 3 – 2 が Yes | TRUE | 判定結果の確信度を想定攻撃者に提示しないようにする |
| 条件 4-1 | データのメタ情報を入手可能 | 質問 4 – 1 が Yes | FALSE | 訓練関連データや推論対象データのメタ情報を想定攻撃者が入手、推定できないようにする |

| | | | | |
|--------|---------------------|--------------|-------|--|
| 条件 4-2 | システムデータの統計情報が入手可能 | 質問 4－2 が Yes | FALSE | 訓練関連データや統計情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-3 | システムデータを 1 個以上入手可能 | 質問 4－3 が Yes | TRUE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-4 | システムデータを千個以上入手可能 | 質問 4－4 が Yes | FALSE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-5 | システムデータを 1 万個以上入手可能 | 質問 4－5 が Yes | FALSE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-6 | 訓練関連データを 1 個以上入手可能 | 質問 4－6 が Yes | TRUE | 訓練関連データを想定攻撃者が入手、推定できないようにする |
| 条件 4-7 | データの正解ラベルが入手可能 | 質問 4－7 が Yes | FALSE | システムの詳細な仕様を公開せず、想定攻撃者に真の正解ラベル |

| | | | | (Ground Truth) を推測されないようにする |
|--------|--------------------|--------------|-------|--|
| 条件 5-1 | 転移性を利用したモデルの開発 | 質問 5－1 が Yes | TRUE | 信頼できる入手先から得られるモデルのみをシステムに組み込む。 あるいは転移性を用いないようにする。 |
| 条件 6-1 | データに対する損失が入手可能 | 質問 6－1 が Yes | FALSE | 損失情報を想定攻撃者が入手できないようにする |
| 条件 6-2 | データに対する勾配が入手可能 | 質問 6－2 が Yes | FALSE | 勾配情報を想定攻撃者が入手できないようにする |
| 条件 6-3 | 訓練済みモデル自体を入手可能 | 質問 6－3 が Yes | FALSE | 訓練済みモデルを管理し、外部に流出しないようにする |
| 条件 7-1 | 類似データを 1 個以上入手可能 | 質問 7－1 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-2 | 類似データを 100 個以上入手可能 | 質問 7－2 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-3 | 類似データを 千個以上入手可能 | 質問 7－3 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 7-4 | 類似データを 1 万個以上入手可能 | 質問 7－4 が Yes | TRUE | システムの利用目的や訓練関連データの詳細な仕様を想定攻撃者が入手、推定できないようにする |
| 条件 8-1 | 扱うデータが テーブルデータ | 質問 8－1 が Yes | FALSE | テーブルデータを扱う AI システムなのか画像データを扱う AI システム |

| | | | | |
|--|--|--|--|-----------------------|
| | | | | ムなのか、利用形態が適切であるかを確認する |
|--|--|--|--|-----------------------|

(iii) 成立した攻撃シナリオ（アタックツリーより判定）

回避攻撃（敵対的サンプル）：A1, A2, A3, A4

ポイズニング攻撃：P2

モデル抽出攻撃：X2

モデルインバージョン攻撃：I1

メンバシップ推測攻撃：M2, M3, M5, M8

(iv) 分析結果

回避攻撃（敵対的サンプル）については全シナリオが成立した。分析者も自身でモデルを任意回数実行可能で、かつ、推論結果を入手可能であるためと考えられる。P2は構築の際にモデルを流用している際に起こりうる攻撃、X2は訓練関連データを入手可能な際に起こりうる攻撃である。他の攻撃シナリオもいくつか実施可能であることが示唆されている。メンバシップ推測攻撃は確信度が入手できるため多くのシナリオが実施可能と判定された。

4. 想定攻撃者＝録画される人々（店舗内の人）

(i) 質問への回答

| 質問番号 | 質問 | 回答 |
|------|---|-----|
| 1-1A | 想定攻撃者の意思で訓練処理を行うことができる AI システムにおいて、自身の意図した数種類のデータで想定攻撃者が訓練処理を実行することができますか？ | - |
| 1-1B | 想定攻撃者でない人が訓練処理を行う AI システム、あるいは自動で訓練処理を行う AI システムにおいて、想定攻撃者が意図したデータを訓練データに混入できますか？ | No |
| 2-1A | 想定攻撃者が準備した推論対象データ 1 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-2A | 【2-1A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 1,000 個以上に対して推論処理を実行することができますか？ | Yes |
| 2-3A | 【2-2A】が Yes であった場合、想定攻撃者は【2-1A】において、意図したデータ 10,000 個以上に対して推論処理を実行す | Yes |

| | | |
|------|--|----|
| | ることができますか？ | |
| 2-4A | 【2-3A】が Yes であった場合、想定攻撃者は 【2-1A】において、意図したデータ 1,000,000 個以上に対して推論処理を実行することができますか？ | No |
| 2-1B | 想定攻撃者が推論対象データを 1 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-2B | 【2-1B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 1,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-3B | 【2-2B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 10,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 2-4B | 【2-3B】が Yes であった場合、想定攻撃者は 【2-1B】において、推論対象データを 1,000,000 個以上置き換えたり、混入したりすることで、推論処理に通すことができますか？ | - |
| 3-1 | 想定攻撃者にモデルの判定結果を提示しますか？ あるいは、想定攻撃者は判定結果を類推することができますか？ | No |
| 3-2 | モデルの確信度を一部でも想定攻撃者に提示しますか？ | No |
| 4-1 | 推論対象データや訓練データを作成できるフォーマット情報（データ形式、サイズの情報、画像の種類、データの次元等のメタ情報）を想定攻撃者は知ることができますか？ | No |
| 4-2 | AI システムに用いられた訓練データの統計情報を想定攻撃者は知ることができますか？ | No |
| 4-3 | AI システムへの入力データそのもの（訓練データ or バリデーションデータ or テストデータ or 推論対象データ）を想定攻撃者が、1 個以上入手できますか？ | No |
| 4-4 | 【4-3】が Yes であった場合、AI システムへの入力データそのもの（訓練データ or テストデータ or 推論対象データ）を想定攻撃者が、1,000 個程度以上入手できますか？ | No |
| 4-5 | 【4-4】が Yes であった場合、AI システムへの入力データそのもの（訓練データ or バリデーションデータ or テストデータ or 推論対象データ）を想定攻撃者が、10,000 個程度以上入手できますか？ | No |
| 4-6 | 訓練関連データに関して、想定攻撃者が 1 個以上入手できますか？ | No |

| | | |
|-----|---|-----|
| 4-7 | AI システムへ入力したデータに対する正解ラベルを想定攻撲者が入手できますか？ | No |
| 5-1 | 外部や内部から入手した訓練済みモデルを一部でも流用し、転移性を利用して AI を構築していますか？ | Yes |
| 6-1 | 推論処理の際に入力された推論対象データに対する損失の情報を想定攻撃者が知ることができますか？ | No |
| 6-2 | 推論処理の際に入力された推論対象データに対する勾配の情報を想定攻撃者が知ることができますか？ | No |
| 6-3 | 訓練済みモデルそのものを想定攻撃者が何らかの方法で入手することができますか？ | No |
| 7-1 | AI システムの仕様として想定される入力データと類似のデータ（ほぼ同じ用途、ほぼ同じ種類・ジャンルのデータ）を、想定攻撃者が 1 個以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-2 | 【7-1】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 100 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-3 | 【7-2】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 1,000 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 7-4 | 【7-3】が Yes であった場合、AI システムの仕様として想定するデータと類似のデータを、想定攻撃者が 10,000 個程度以上、何らかの手段で準備・入手できますか？ | Yes |
| 8-1 | システムで扱っているデータはテーブルデータですか？ | No |

(ii) 条件合致性判定

| 攻撃実施可能条件 | 内容 | TRUE になる条件 | 判定結果 (TRUE/FALSE) | FALSE にする対策案 |
|----------|------------------|--------------------------------|-------------------|--------------------------|
| 条件 1-1 | 訓練処理の自由な実行が可能 | 質問 1 - 1 A または質問 1 - 1 B が Yes | FALSE | 想定攻撃者が訓練処理を実行できないようにする |
| 条件 2-1 | 推論処理を 1 個以上で実行可能 | 質問 2 - 1 A または質問 2 - 1 B が Yes | TRUE | 想定攻撃者が推論処理を実行できないように設定する |
| 条件 2-2 | 推論処理を千個以上で実行可能 | 質問 2 - 2 A または質問 2 - 2 B が Yes | TRUE | 想定攻撃者がデータ 1,000 個以上に対して推 |

| | | | | |
|--------|---------------------|--------------------------------|-------|--|
| | | | | 論処理を実行できないよう設定する |
| 条件 2-3 | 推論処理を 1 万個以上で実行可能 | 質問 2 – 3 A または質問 2 – 3 B が Yes | TRUE | 想定攻撃者がデータ 10,000 個以上に対して推論処理を実行できないように設定する |
| 条件 2-4 | 推論処理を 100 万個以上で実行可能 | 質問 2 – 4 A または質問 2 – 4 B が Yes | FALSE | 想定攻撃者がデータ 1,000,000 個以上に対して推論処理を実行できないように設定する |
| 条件 3-1 | 推論結果を入手可能 | 質問 3 – 1、または質問 3 – 2 が Yes | FALSE | 判定結果を想定攻撃者に提示しないようにする |
| 条件 3-2 | 確信度を入手可能 | 質問 3 – 2 が Yes | FALSE | 判定結果の確信度を想定攻撃者に提示しないようにする |
| 条件 4-1 | データのメタ情報を入手可能 | 質問 4 – 1 が Yes | FALSE | 訓練関連データや推論対象データのメタ情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-2 | システムデータの統計情報を入手可能 | 質問 4 – 2 が Yes | FALSE | 訓練関連データや統計情報を想定攻撃者が入手、推定できないようにする |
| 条件 4-3 | システムデータを 1 個以上入手可能 | 質問 4 – 3 が Yes | FALSE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撃者が入手、推定できないようにする。 |
| 条件 4-4 | システムデータを千個以上入手可能 | 質問 4 – 4 が Yes | FALSE | 訓練関連データや過去に使用した推論対象データを想定攻撃者が入手できないようにする。 |

| | | | | |
|--------|---------------------|--------------|-------|--|
| | | | | また、AI システムのタスクや入力データのメタ情報を想定攻撲者が入手、推定できないようにする。 |
| 条件 4-5 | システムデータを 1 万個以上入手可能 | 質問 4-5 が Yes | FALSE | 訓練関連データや過去に使用した推論対象データを想定攻撲者が入手できないようにする。 また、AI システムのタスクや入力データのメタ情報を想定攻撲者が入手、推定できないようにする。 |
| 条件 4-6 | 訓練関連データを 1 個以上入手可能 | 質問 4-6 が Yes | FALSE | 訓練関連データを想定攻撲者が入手、推定できないようにする |
| 条件 4-7 | データの正解ラベルが入手可能 | 質問 4-7 が Yes | FALSE | システムの詳細な仕様を公開せず、想定攻撲者に真の正解ラベル (Ground Truth) を推測されないようにする |
| 条件 5-1 | 転移性を利用したモデルの開発 | 質問 5-1 が Yes | TRUE | 信頼できる入手先から得られるモデルのみをシステムに組み込む。 あるいは転移性を用いないようにする。 |
| 条件 6-1 | データに対する損失が入手可能 | 質問 6-1 が Yes | FALSE | 損失情報を想定攻撲者が入手できないようにする |
| 条件 6-2 | データに対する勾配が入手可能 | 質問 6-2 が Yes | FALSE | 勾配情報を想定攻撲者が入手できないようにする |
| 条件 6-3 | 訓練済みモデル自体を入手可能 | 質問 6-3 が Yes | FALSE | 訓練済みモデルを管理し、外部に流出しないようにする |

| | | | | |
|--------|---------------------------|-----------------|-------|--|
| 条件 7-1 | 類似データを 1 個以上入手 可能 | 質問 7－1 が Yes | TRUE | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 7-2 | 類似データを 100 個以上入 手可能 | 質問 7－2 が Yes | TRUE | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 7-3 | 類似データを 千個以上入手 可能 | 質問 7－3 が Yes | TRUE | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 7-4 | 類似データを 1 万個以上入 手可能 | 質問 7－4 が Yes | TRUE | システムの利用目的や訓 練関連データの詳細な仕 様を想定攻撃者が入手、 推定できないようにする |
| 条件 8-1 | 扱うデータが テーブルデー タ | 質問 8－1 が Yes | FALSE | テーブルデータを扱う AI システムなのか画像 データを扱う AI システ ムなのか、利用形態が適 切であるかを確認する |

(iii) 成立した攻撃シナリオ（アタックツリーより判定）

回避攻撃（敵対的サンプル）：なし

ポイズニング攻撃：P2

モデル抽出攻撃：なし

モデルインバージョン攻撃：なし

メンバシップ推測攻撃：なし

(iv) 分析結果

P2 のみ成立となった。P2 はモデルの構造に起因しているので実施可能となった。P2 は流用したモデルがはじめから汚染させていた際に起こりうる攻撃であり、想定攻撃者とした第三者が実際に汚染できるわけではない。結果的に第三者による攻撃は考えにくい。

II-9. まとめ

本ガイドラインでは、AI 開発者が自身で脅威分析を行うための分析技術の構築方法として、選択回答式 AI セキュリティリスク問診（AI リスク問診）を紹介した。また、AI リスク問診の構築例、及び、試行結果を例示した。本技術によって得られたアタックツリーの成立・不成立の情報より、成立したアタックツリーに対応する攻撃シナリオは実施可能になると判断できる。成立したアタックツリーに対応する攻撃を実施困難にするためには、アタックツリーを不成立にするための条件を考察し、不成立にするための仕様変更を行うことで実施する。本技術は AI 開発者が自身で分析して対応策を考えるための支援技術に相当し、自身での対応策検討を行ったり、本技術の分析結果を参考資料として AI セキュリティ専門家へ相談する際に利用したりすることもできる。なお、本技術によって得られた対応案を何らかの状況、条件で実施できない時には AI セキュリティ専門家に相談して専用の対策を検討して頂きたい。本技術の今後についてはさらなる攻撃シナリオへの対応が想定される。本技術を活用し、機械学習システムのセキュリティ強化を検討して頂きたい。

II-10. 参考文献

- [II-1] 総務省 AI ネットワーク社会推進会議, “AI 利活用ガイドライン～AI 利活用のためのプラクティカルリファレンス～”
https://www.soumu.go.jp/main_content/000637097.pdf
- [II-2] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, “Robust Physical-World Attacks on Deep Learning Models”, CVPR 2018.
- [II-3] European Union Agency for Cybersecurity (ENISA), “Artificial Intelligence Cybersecurity Challenges”, 2020.
<https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges>
- [II-4] Microsoft, “AI/ML システムと依存関係の脅威のモデル化”, 2019.
<https://docs.microsoft.com/ja-jp/security/engineering/threat-modeling-aiml>
- [II-5] 矢嶋, 清水, 森川, 大久保, “機械学習システムに潜む AI セキュリティ脆弱性の分析手法に関する一考察”, 2021 年暗号と情報セキュリティシンポジウム
- [II-6] 矢嶋, 及川, 森川, 笠原, 乾, 吉岡, “開発エンジニア向け機械学習セキュリティ脅威分析技術”, 2022 年暗号と情報セキュリティシンポジウム
- [II-7] M. Juuti, S. Szylner, S. Marchal, N. Asokan, “PRADA: Protecting against DNN Model Stealing Attacks”, the 4th IEEE European Symposium on Security and Privacy (EuroS&P 2019)
- [II-8] T. Orekondy, B. Schiele, M. Fritz, “Knockoff Nets: Stealing Functionality of Black-Box Models”, arXiv <https://arxiv.org/abs/1812.02766>
- [II-9] R. Shokri, M. Stronati, C. Song, V. Shmatikov, “Membership Inference Attacks Against Machine Learning Models”, 2017 IEEE Symposium on Security and Privacy (S&P).
- [II-10] S. Yeom, I. Giacomelli, M. Fredrikson, S. Jha, “Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting”, 2018 IEEE 31st Computer Security Foundations Symposium (CSP).
- [II-11] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, M. Backes, “ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models”, The Network and Distributed System Security 2019 (NDSS 2019).
- [II-12] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, Herve Jegou, “White-box vs Black-box: Bayes Optimal Strategies for Membership Inference”, the 36th International Conference on Machine Learning.
- [II-13] Z. Li, Y. Zhang, “Membership Leakage in Label-Only Exposures”, 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS’2021).

機械学習システム セキュリティガイドライン

「付録：攻撃検知技術の概要」

Version 1.02

2022 年 9 月 16 日

機械学習システムセキュリティガイドライン策定委員会

機械学習システムセーフティ・セキュリティワーキンググループ

日本ソフトウェア科学会 機械学習工学研究会



A-1. はじめに

機械学習システムには、判断を誤らせたり情報を奪ったりするような、機械学習特有の攻撃が存在するため、従来のセキュリティ対策だけでなく、機械学習特有の攻撃への対策も必要となる。対策として、攻撃に堅牢なシステムを構築することに加え、攻撃を検知することも重要となる。

機械学習への攻撃を検知する手法は多く提案されているが、特定のデータやタスクに対してのみ有効な場合も多く、決定的な検知手法はまだ確立されていない。Carlini らは敵対的サンプルの検知について既存の代表的な手法を評価し、検知手法を知った上で攻撃に対しては十分な検知性能が得られないことを示している [A-1]。また、Kumar らは機械学習セキュリティの課題の一つに攻撃検知を挙げており、セキュリティ分析者の間で検知に関する知見の共有ができていないことについて言及している [A-2]。したがって現状では、機械学習セキュリティの専門知識が無ければ、適切な検知手法を実装することは容易ではないと言える。

一般的なサイバー攻撃対策においては、攻撃戦略や手法を体系化した MITRE ATT&CK [A-3]などのフレームワークがあり、検知手法を検討する際にも活用されている [A-4]。機械学習システムへの攻撃については、MITRE ATLAS [A-5]において機械学習特有の攻撃戦略（攻撃に必要な情報収集などの偵察活動、攻撃の実行など）が整理されており、検知対象とする攻撃の選定の参考になる。一方で、著者が知る限り、機械学習特有の攻撃を検知する手法を攻撃戦略に沿ってまとめた文献はない。そこで本付録では、攻撃戦略を「偵察活動や攻撃のための下準備」（攻撃の前兆）と、「実際に誤判定を引き起こさせたり情報を奪ったりする攻撃」（攻撃の兆候）の 2 段階とし、それぞれを検知する手法を前兆検知と兆候検知に分けて整理した。また、検知に使用するデータが機械学習システムにおいて取得可能かどうかも検知手法選定時には重要な情報となるため、使用するデータ（訓練データ、訓練済みモデルなど）の観点からも検知手法を整理した。

本付録は、開発者やセキュリティ分析者の検知手法選定を支援することを目的とし、攻撃検知に関する文献を攻撃戦略と使用するデータの観点から整理した。なお、対象とするシステムは画像分類システムとし、対象とする攻撃は回避攻撃、ポイズニング攻撃、モデル抽出攻撃とする（対象範囲は今後拡張していく予定）。

A-2. 攻撃戦略毎の検知手法について

本付録では、攻撃戦略を「偵察活動や攻撃のための下準備」と、「実際に誤判定を引き起こさせたり情報を奪ったりする攻撃」の 2 段階に分けて考える。前者を攻撃の前兆、後者を攻撃の兆候と呼ぶこととし、それぞれの検知を前兆検知と兆候検知とする。本節では回避攻撃、ポイズニング攻撃、モデル抽出攻撃の検知手法を前兆検知と兆候検知という 2 つの検知目的に分けて整理した（整理した一覧表は表 A-1 に示す）。

A-2.1. 前兆検知

A-2.1.1. 回避攻撃の検知

画像分類を行う機械学習システムの場合、攻撃者は入力画像に対して撮動と呼ばれる小さなノイズを

加えることで、機械学習モデルを誤認識させることができる [A-6]。このような振動を加えて誤認識させるようなデータを敵対的サンプルと呼ぶ(敵対的サンプルに関する論文リスト (arXiv) が Nicholas Carlini によってまとめられている [A-7])。

敵対的サンプルの作成方法は攻撃者が有している知識によっても異なる。例えば、攻撃対象のシステムに対する知識を持っていない場合は、システムに対して一連のクエリを実行して敵対的サンプルを作成する方法がある [A-8]。一方、攻撃対象システムに対する知識を有している場合は、攻撃者の手元で攻撃対象モデルを再現し、攻撃対象システムへのクエリを実行せずに敵対的サンプルを作成できる [A-9]。前者のようにクエリ実行を通して敵対的サンプルを作成しようとする活動は、攻撃の前兆として攻撃が成功（敵対的サンプル作成）する前に検知することが望ましい。検知手法について以下に示す。

■ 敵対的サンプルを作成しようとする操作を検知する

攻撃者が敵対的サンプルを作成するためには、複数の類似するデータを入力する可能性が高いため、攻撃者が入力する一連のクエリが正常ユーザとは異なる分布になることを利用する検知手法が提案されている [A-10, 11]。また、このような攻撃はクエリ数が正常ユーザと比べて多くなる場合や、出力ラベルに偏りが出る場合が多いため、単位時間当たりのクエリ頻度や出力ラベルの分布をモニタリングする等、単純な方法で検知できる可能性もある。

A-2.1.2. ポイズニング攻撃の検知

ポイズニング攻撃には、訓練データを汚染することによりモデルの推論精度を意図的に劣化させる攻撃 [A-12] と、訓練データにバックドアデータを仕込み、特定の入力データを攻撃者が意図したクラスに誤分類させるようなバックドア攻撃がある [A-13]。訓練データの準備やモデル作成のアウトソーシング [A-14]、信頼できない Web サイトからのデータ収集、連合学習 [A-15] や転移学習 [A-16] によってデータやモデルが汚染される恐れがある [A-17]。

これらは攻撃の前兆ととらえることができ、テスト段階、もしくはそれよりも早期に検知することが望ましい。運用中の入力データを訓練データに加えて再学習するような場合も、入力されたデータが汚染されていないかを確認してから再学習する必要がある。検知手法について以下に示す。

■ 精度劣化を引き起こすような訓練データセットの汚染を検知する

訓練用に収集したデータに、精度劣化を引き起こすようなデータが含まれていないかを検知する手法が提案されている [A-18, 19]。

■ 訓練データセットのバックドアを検知する

訓練データセットに含まれるバックドアデータを検知する手法が提案されている [A-20, 21]。文献 [A-17] はバックドア攻撃やその対策に関する包括的なレビューであり、検知手法に関しても整理されているため参照されたい。

■ 訓練済みモデルが汚染されていることを検知する

外部から取得した訓練済みモデル等が汚染されていないかを検知する手法が提案されている [A-22, 23]。

A-2.2.兆候検知

A-2.2.1.回避攻撃の検知

■ 敵対的サンプルの入力を検知する

前述の通り、攻撃者が攻撃対象システムに対する知識を有している場合などは、何らかの方法で敵対的サンプルを作成し、システムに入力してくる可能性がある。このような攻撃を検知するためには、入力データ、モデルの出力結果、中間層出力のデータを分析することが多い [A-24-32]。ただし、敵対的サンプルの入力検知は一般的に難しいタスクと言われており [A-1]、あらゆるシステムに有効な検知手法はないため、可能な限り複数の検知手法を適用することを推奨する。

敵対的サンプルの検知手法に関する包括的なレビューは文献 [A-33]を参照すること。

A-2.2.2.ポイズニング攻撃の検知

■ 入力データにバックドアのトリガーが含まれるかを検知する

運用時にトリガーが含まれる画像が入力されていないかを検知する手法が提案されている [A-34]。

A-2.2.3.モデル抽出攻撃の検知

モデル抽出攻撃は、システムへの複数クエリとともにモデルの入出力のペアを取得し、取得した情報をもとにシステムで使用されているモデルと似たふるまいをするモデルを作成する攻撃である。攻撃者にとってモデル抽出自体が目的である場合もあれば、他の攻撃に用いるためにモデル抽出攻撃を行う場合もある [A-35]（後者の場合は攻撃の前兆とも言える）。モデル抽出攻撃を検知する手法を以下に示す。

■ モデル抽出攻撃を目的とした異常なクエリを検知する

モデル抽出攻撃を実施する際は正常ユーザとは異なるログになる可能性が高いため、それを利用して検知する手法が提案されている [A-36-39]。また、このような攻撃はクエリ数が正常ユーザと比べて多くの場合があるため、単位時間当たりのクエリ頻度をモニタリングする等、単純な方法で検知できる可能性もある。

A-3. 検知に使用するデータについて

「検知手法を実装する際に使用するデータ」の観点から既存の検知手法を整理する。ここでいうデータは、訓練データ、訓練済みモデル、運用中の入力データ、モデルの出力データの4つを考える。また、これらのデータは、検知手法によって、「検知・分析の対象となるデータ」と、「攻撃の痕跡はないが、検知・分析する際に必要なデータ」に分けることができる。前者は日時やアカウント等と紐づけて管理する必要があり（前者の例：敵対的サンプルの撮動の痕跡が残っている可能性がある入力画像）、後者は検知手法を実装する際に必要となるため適切に管理する必要がある（後者の例：攻撃判定に用いる閾値を算出するために用いる訓練データ）。上記の観点から検知手法を整理したものを表 A-1 に示す。なお、出力データはモデルが出力する確信度や中間層出力等を指し、検知手法によって異なるため、詳細については各文献を参照されたい。

表 A-1. 検知に使用するデータについて

●：検知・分析の対象となるデータ

□：攻撃の痕跡はないが、検知・分析する際に必要なデータ

| 検知目的 | 検知対象攻撃 | 検知手法 | 開発 | | 運用 | |
|------|----------------|---------------------------|-------|--------|-------|-------|
| | | | 訓練データ | 訓練済モデル | 入力データ | 出力データ |
| 前兆検知 | 敵対的サンプル作成の検知 | Chen et al. [A-10] | □ | | ● | |
| | | Li et al. [A-11] | □ | | ● | |
| | データ汚染（精度劣化）の検知 | Müller et al. [A-18] | ● | □ | | |
| | | Tavallali et al. [A-19] | ● | | | |
| | バックドアの検知 | Chen et al. [A-20] | ● | □ | | |
| | | Hayase et al. [A-21] | ● | □ | | |
| | | Dong et al. [A-22] | | ● | | |
| | | Huster et al. [A-23] | | ● | | |
| 兆候検知 | 敵対的サンプル入力の検知 | Hendrycks et al. [A-24] | □ | | ● | |
| | | Meng et al. [A-25] | □ | | ● | |
| | | Grosse et al. [A-26] | □ | | ● | |
| | | Gong et al. [A-27] | □ | □ | ● | |
| | | Lu et al. [A-28] | □ | □ | | ● |
| | | Feinman et al. [A-29] | □ | □ | | ● |
| | | Aigrainet al. [A-30] | □ | □ | | ● |
| | | Xu et al. [A-31] | □ | □ | ● | ● |
| | | Monteiro et al. [A-32] | □ | □ | ● | ● |
| | バックドアトリガーの検知 | Kiourti et al. [A-34] | □ | □ | ● | ● |
| | モデル抽出攻撃の検知 | Juuti et al. [A-36] | | | ● | |
| | | Pal et al. [A-37] | □ | | ● | |
| | | Atli et al. [A-38] | □ | | ● | |
| | | Sadeghzadeh et al. [A-39] | □ | | ● | |

A-4. まとめ

本付録では、開発者やセキュリティ分析者の検知手法選定を支援することを目的とし、回避攻撃、ポイズニング攻撃、モデル抽出攻撃の検知に関する文献を攻撃戦略と使用するデータの観点から整理した。攻撃戦略に従って検知手法を整理することによって、一連の攻撃をその前兆と兆候で多段階に検知する場合や、可能な限り早い段階で攻撃を検知したい場合等の参考になると考えられる。また、使用するデータの観点から検知手法を整理することによって、開発するシステムにおいて取得するデータに制約がある場合でも（例：入力データを取得できない）、取得できるデータのみを用いた検知手法を選定できると考えられる。既存のレビュー文献 [A-17, 33] 等とも合わせて検知手法選定時に参照されたい。

A-5. 参考文献

- [A-1] N. Carlini , D. Wagner, “Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods,” In Workshop on Artificial Intelligence and Security, 2017.
- [A-2] R. S. S. Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann , S. Xia, “Adversarial Machine Learning -- Industry Perspectives,” In IEEE Security and Privacy Workshops, 2020.
- [A-3] “ATT&CK,” MITRE, [オンライン]. Available: <https://attack.mitre.org/>.
- [A-4] B. E. Strom, J. A. Battaglia, M. S. Kemmerer, W. Kupersanin, D. P. Miller, C. Wampler, S. M. Whitley , a. R. D. Wolf, “Finding Cyber Threats with ATT&CK™-Based Analytics,” The MITRE Corporation, Bedford, MA, Technical Report No. MTR170202, 2017.
- [A-5] “ATLAS,” MITRE, [オンライン]. Available: <https://atlas.mitre.org/>.
- [A-6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow , R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.
- [A-7] N. Carlini, “A Complete List of All (arXiv) Adversarial Example Papers,” <https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html>, 2019.
- [A-8] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi , C.-J. Hsieh, “ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models,” In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 15–26, 2017.
- [A-9] I. J. Goodfellow, J. Shlens , C. Szegedy, “Explaining and Harnessing Adversarial Examples,” In International Conference on Learning Representations, 2015.
- [A-10] S. Chen, N. Carlini , D. Wagner, “Stateful Detection of Black-Box Adversarial Attacks,” In Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence, pp.30-39, 2019.
- [A-11] H. Li, S. Shan, E. Wenger, J. Zhang, H. Zheng , B. Y. Zhao, “Blacklight: Defending Black-Box Adversarial Attacks on Deep Neural Networks,” arXiv preprint arXiv:2006.14042, 2020.
- [A-12] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru , B. Li, “Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning,” In IEEE Security and Privacy, 2018.
- [A-13] T. Gu, B. Dolan-Gavitt , S. Garg, “BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain,” In Proceedings of Machine Learning and Computer Security Workshop, 2017.
- [A-14] Y. Chen, X. Gong, Q. Wang, X. Di , H. Huang, “Backdoor Attacks and Defenses for Deep Neural Networks in Outsourced Cloud Environments,” IEEE Network, vol. 34, no. 5, pp. 141–147, 2020.
- [A-15] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin , V. Shmatikov, “How To Backdoor Federated Learning,” In International Conference on Artificial Intelligence and Statistics, 2020.

- [A-16] Y. Ji, Z. Liu, X. Hu, P. Wang , Y. Zhang, “Programmable Neural Network Trojan for Pre-Trained Feature Extractor,” arXiv preprint arXiv:1901.07766, 2019.
- [A-17] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal , H. Kim, “Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review,” arXiv preprint arXiv:2007.10760, 2020.
- [A-18] N. M. Müller, S. Roschmann , K. Böttlinger, “Defending Against Adversarial Denial-of-Service Data Poisoning Attacks,” arXiv preprint arXiv:2104.06744, 2021.
- [A-19] P. Tavallali, V. Behzadan, P. Tavallali , M. Singhal, “Adversarial Poisoning Attacks and Defense for General Multi-Class Models Based On Synthetic Reduced Nearest Neighbors,” arXiv preprint arXiv:2102.05867, 2021.
- [A-20] J. Chen, X. Zhang, R. Zhang, C. Wang , L. Liu, “De-Pois: An Attack-Agnostic Defense against Data Poisoning Attacks,” In IEEE Transactions on Information Forensics and Security, 2021.
- [A-21] J. Hayase, W. Kong, R. Somani , S. Oh, “SPECTRE: Defending Against Backdoor Attacks Using Robust Statistics,” In Proceedings of the 38th International Conference on Machine Learning, 2021.
- [A-22] Y. Dong, X. Yang, Z. Deng, T. Pang, Z. Xiao, H. Su , J. Zhu, “Black-box Detection of Backdoor Attacks with Limited Information and Data,” In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
- [A-23] T. Huster , E. Ekwedike, “TOP: Backdoor Detection in Neural Networks via Transferability of Perturbation,” arXiv preprint arXiv:2103.10274, 2021.
- [A-24] D. Hendrycks , K. Gimpel, “Early Methods for Detecting Adversarial Images,” In Proceedings of the International Conference on Learning Representations, 2017.
- [A-25] D. Meng , H. Chen, “MagNet: a Two-Pronged Defense against Adversarial Examples,” In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 135–147, 2017.
- [A-26] K. Grosse, P. Manoharan, N. Papernot, M. Backes , P. McDaniel, “On the (Statistical) Detection of Adversarial Examples,” arXiv preprint arXiv:1702.06280, 2017.
- [A-27] Z. Gong, W. Wang , W.-S. Ku, “Adversarial and Clean Data Are Not Twins,” arXiv preprint arXiv:1704.04960, 2017.
- [A-28] J. Lu, T. Issaranon , D. Forsyth, “SafetyNet: Detecting and Rejecting Adversarial Examples Robustly,” In Proceedings of the IEEE International Conference on Computer Vision, pp. 446–454, 2017.
- [A-29] R. Feinman, R. R. Curtin, S. Shintre , A. B. Gardner, “Detecting Adversarial Samples from Artifacts,” arXiv preprint arXiv:1703.00410, 2017.
- [A-30] J. Aigrain , M. Detyniecki, “Detecting Adversarial Examples and Other Misclassifications in Neural Networks by Introspection,” In Proceedings of the 35th International Conference on Machine Learning, pp. 7167–7177, 2019.

- [A-31] W. Xu , Y. Q. David Evans, “Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks,” In Proceedings of Network and Distributed System Security Symposium, 2018.
- [A-32] J. Monteiro, I. Albuquerque, Z. Akhtar , T. H. Falk, “Generalizable Adversarial Examples Detection Based on Bi-model Decision Mismatch,” In 2019 IEEE International Conference on Systems, Man and Cybernetics, pp. 2839–2844, 2019.
- [A-33] A. Aldahdooh, W. Hamidouche, S. A. Fezza , O. Deforges, “Adversarial Example Detection for DNN Models: A Review and Experimental Comparison,” arXiv preprint arXiv:2105.00203, 2021.
- [A-34] P. Kiourtis, W. Li, A. Roy, K. Sikka , S. Jha, “MISA: Online Defense of Trojaned Models using Misattributions,” arXiv preprint arXiv:2103.15918, 2021.
- [A-35] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik , A. Swami, “Practical Black-Box Attacks against Machine Learning,” In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp. 506–519, 2017.
- [A-36] M. Juuti, S. Szylner, S. Marchal , N. Asokan, “PRADA: Protecting against DNN Model Stealing Attacks,” In IEEE European Symposium on Security & Privacy, pp. 512–527, 2019.
- [A-37] S. Pal, Y. Gupta, A. Kanade , S. Shevade, “Stateful Detection of Model Extraction Attacks,” arXiv preprint arXiv:2107.05166, 2021.
- [A-38] B. G. Atli, S. Szylner, M. Juuti, S. Marchal , N. Asokan, “Extraction of Complex DNN Models: Real Threat or Boogeyman?,” In International Workshop on Engineering Dependable and Secure Machine Learning Systems. Springer, pp. 42–57, 2020.
- [A-39] A. M. Sadeghzadeh, F. Dehghan, A. M. Sobhanian , R. Jalili, “HODA: Hardness-Oriented Detection of Model Extraction Attacks,” arXiv preprint arXiv:2106.11424, 2021.

機械学習システムセキュリティガイドライン策定委員会メンバーリスト

市原 大暉 (株式会社 NTT データ)
及川 孝徳 (富士通株式会社)
大久保 隆夫 (情報セキュリティ大学院大学)
笠原 史禎 (富士通株式会社)
金子 朋子 (国立情報学研究所)
久連石 圭 (株式会社 東芝)
田口 研治 (国立情報学研究所)
林 昌純 (法政大学)
森川 郁也 (富士通株式会社)
矢嶋 純 (富士通株式会社)
吉岡 信和 (早稲田大学)
(敬称略・五十音順)