

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 1.03a  
2022年12月26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

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

## 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」

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-61
II-8.1.3. 性別・年齢推定 AI .....	II-66
II-9.まとめ .....	II-89
II-10.参考文献 .....	II-90
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 .....	II-91

## 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。

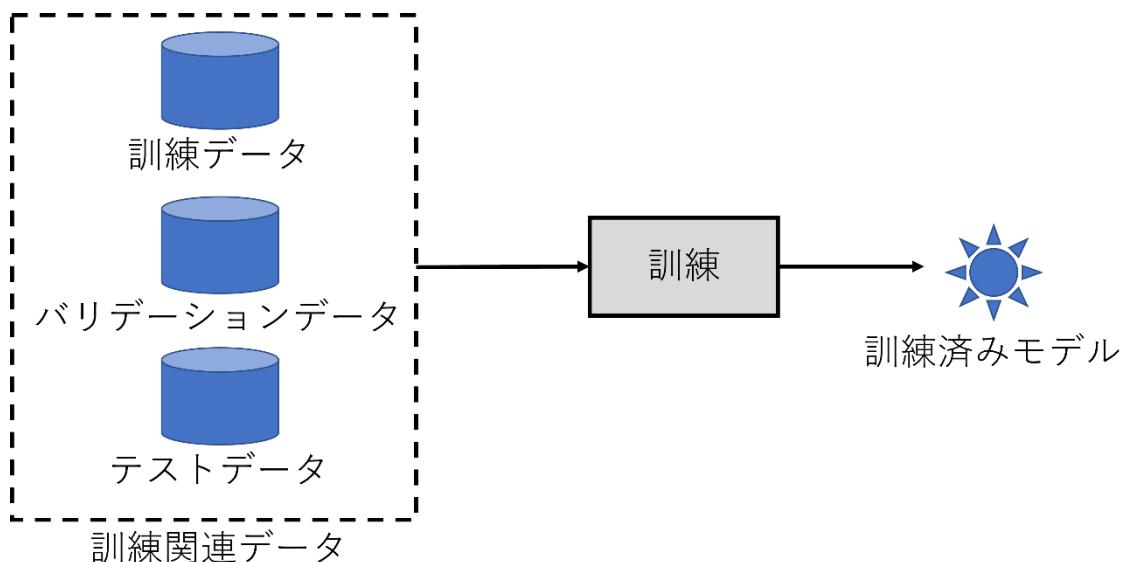


図 II- 1. 機械学習システムの訓練パイプライン

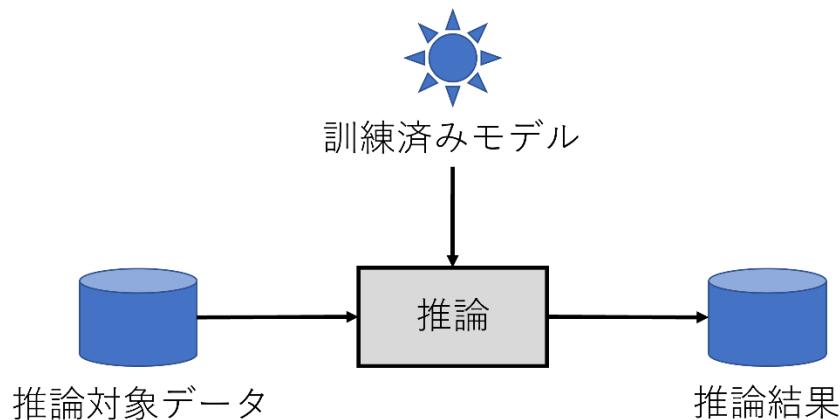


図 II- 2. 機械学習システムの推論パイプライン

## 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。

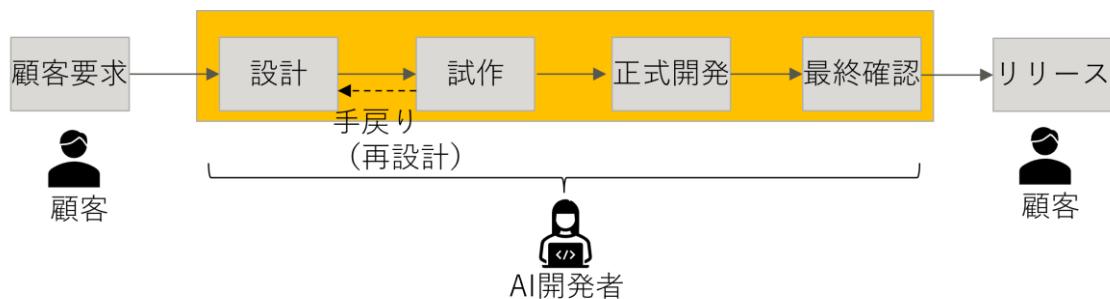


図 II- 3. 機械学習システムの機械学習処理部における一般的な開発プロセス

## 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な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際に攻撃が成功する。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する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。

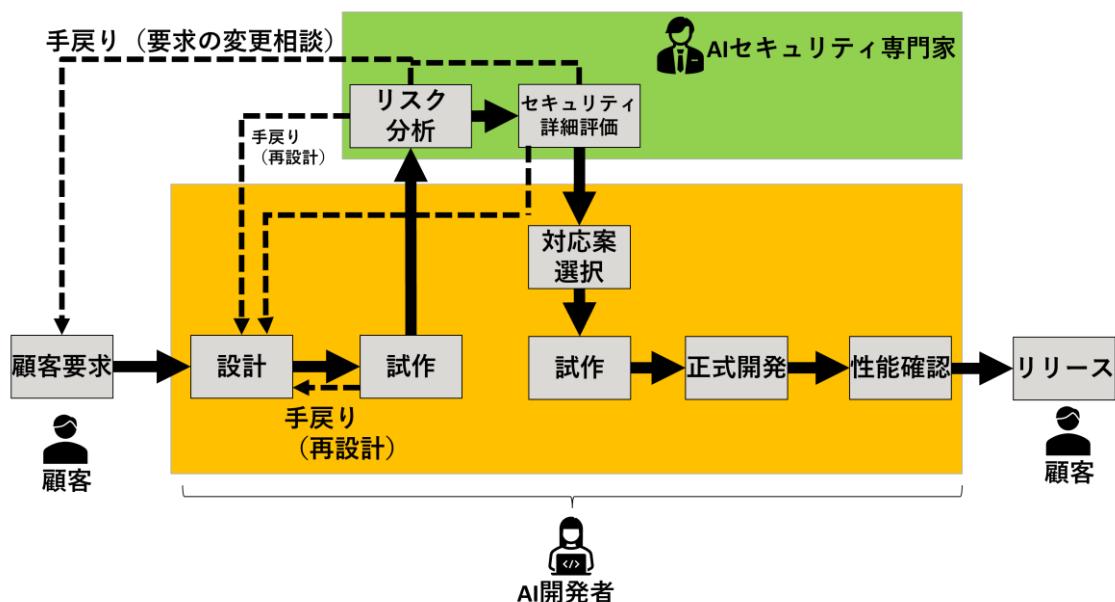


図 II-4. 機械学習システム特有の攻撃に対するセキュリティに対応した  
機械学習処理部の開発プロセス

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。

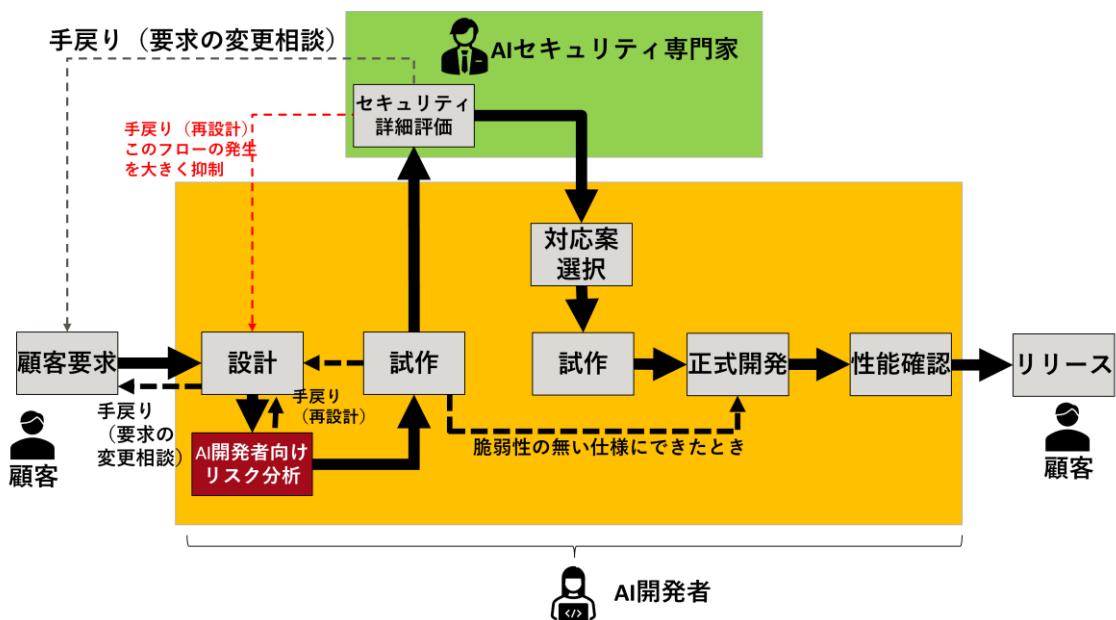


図 II- 5. セキュリティ対策による手戻りを抑制した  
機械学習処理部の開発プロセス（目指すべき形）

## 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することができる。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。

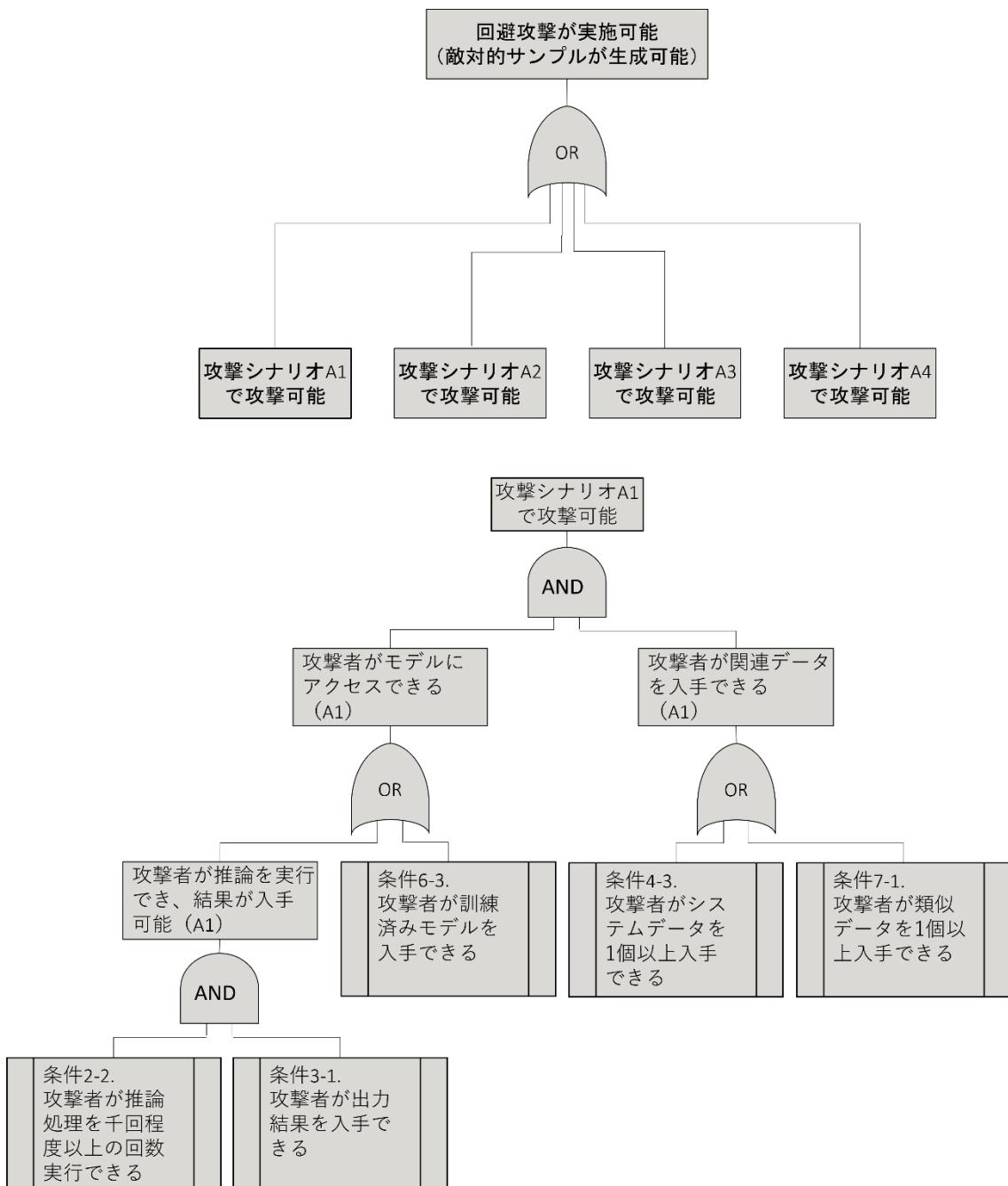


図 II- 6. 構成したアタックツリー（の一部）の例

#### 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ー
- 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。

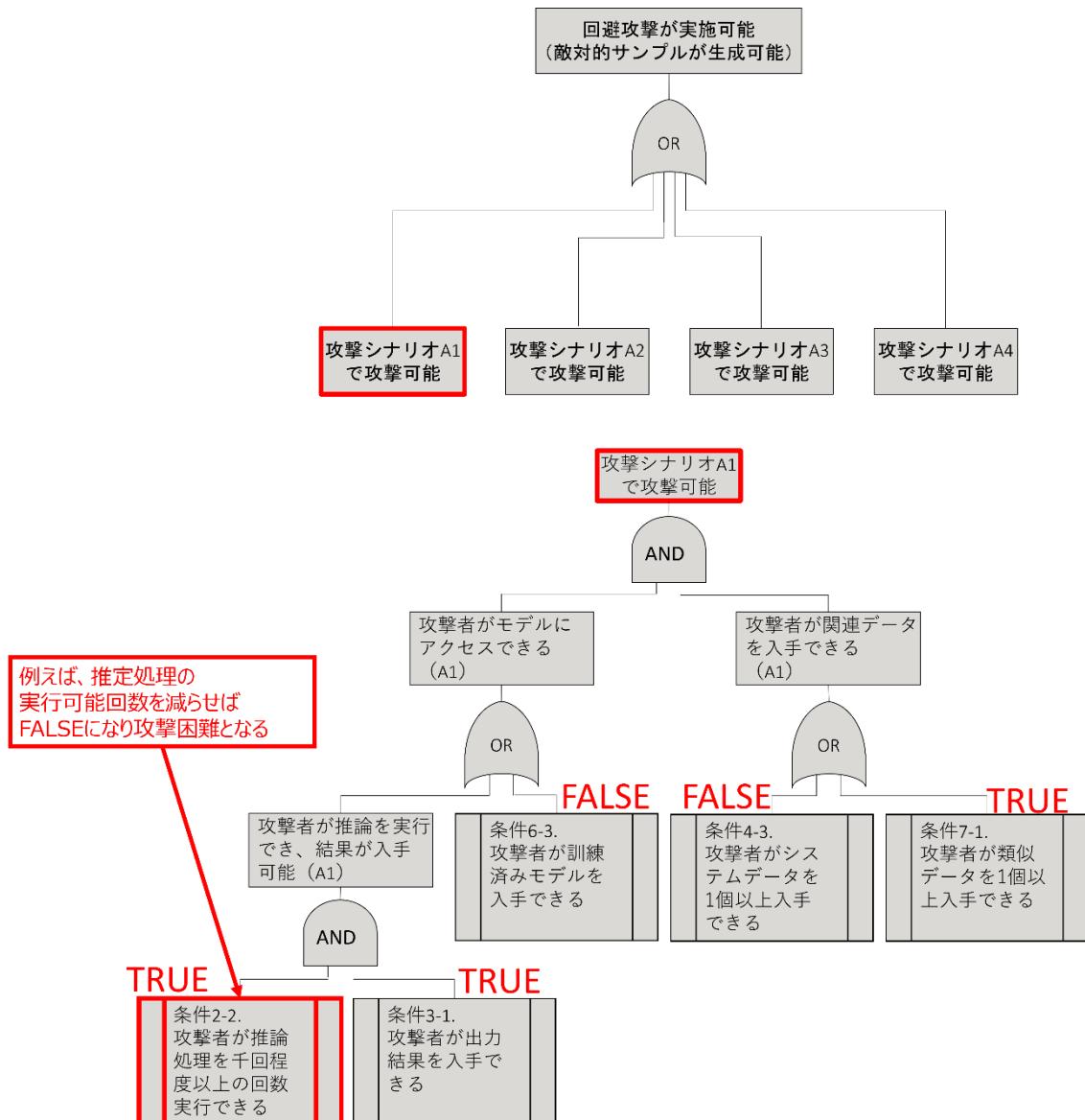


図 II- 7. 対策の検討例

## 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。

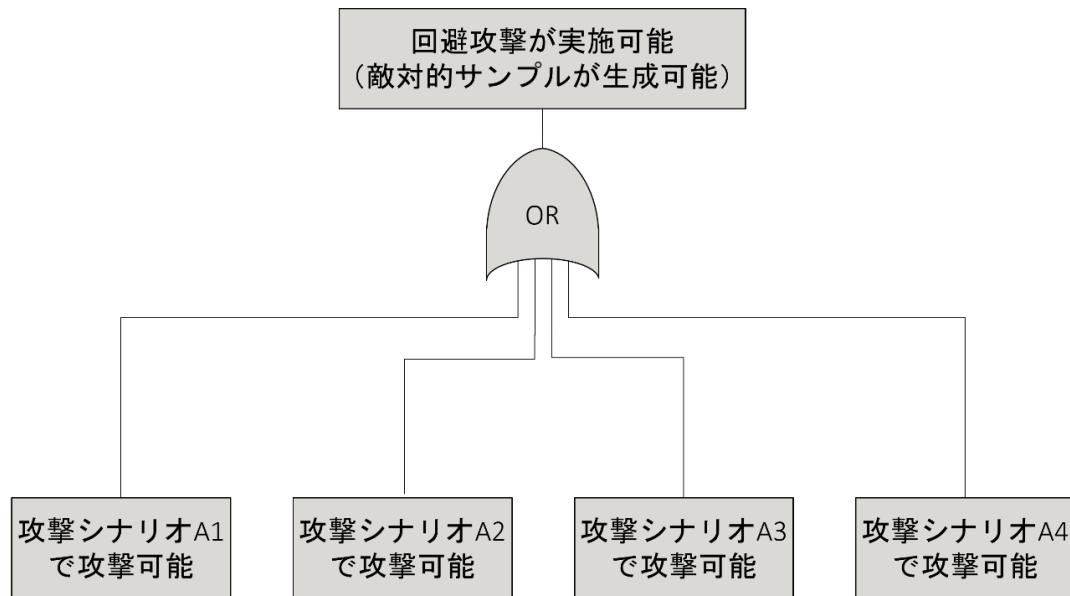


図 II- 8. 回避攻撃（敵対的サンプル）のアタックツリー（上位部分）

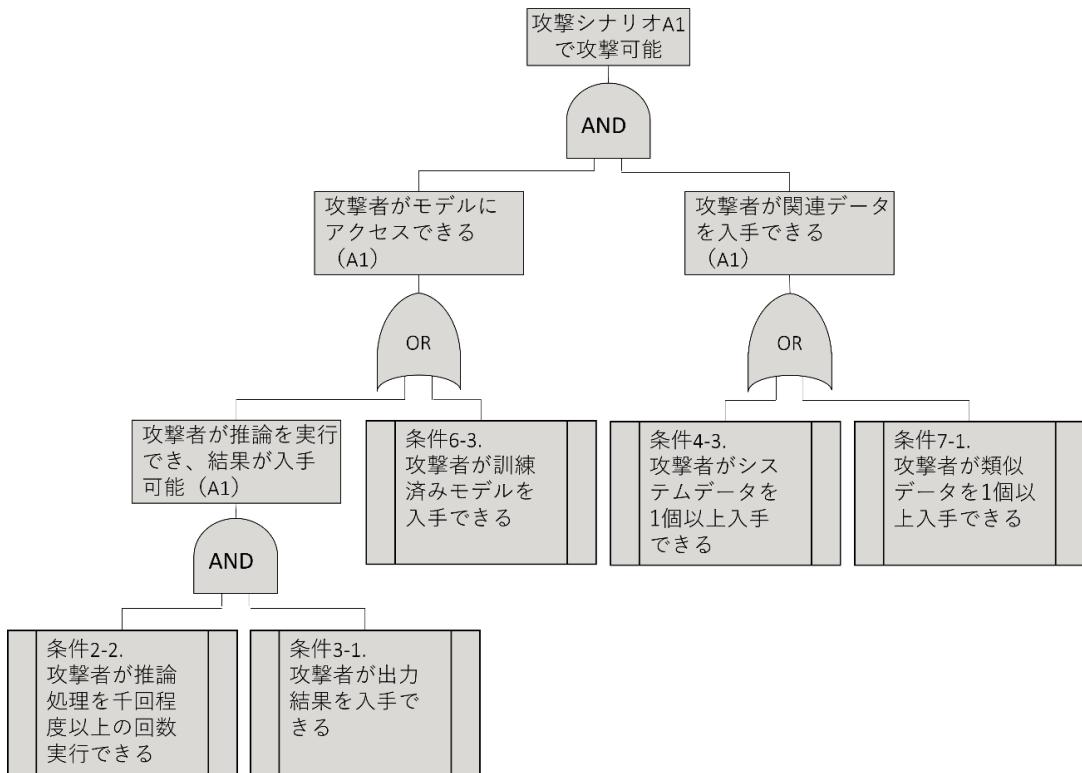


図 II- 9. 回避攻撃（敵対的サンプル）の攻撃シナリオ A1 の  
アタックツリーと攻撃実施可能条件

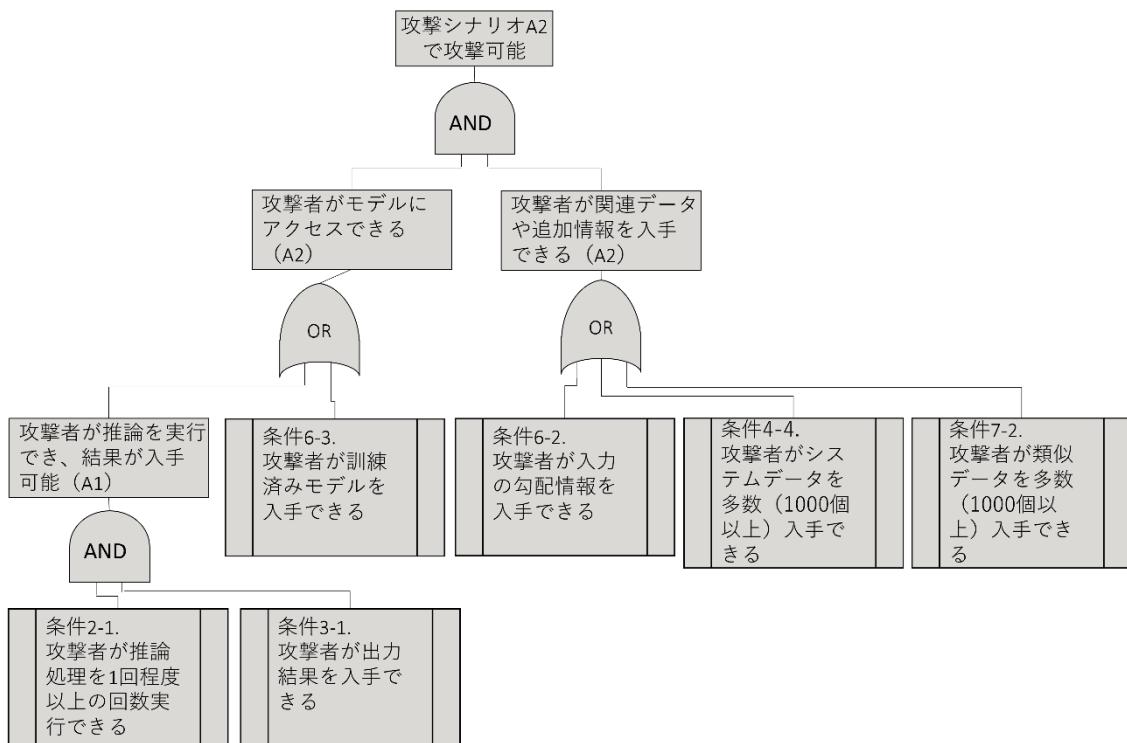


図 II- 10. 回避攻撃（敵対的サンプル）の攻撃シナリオ A2 のアタックツリーと攻撃実施可能条件

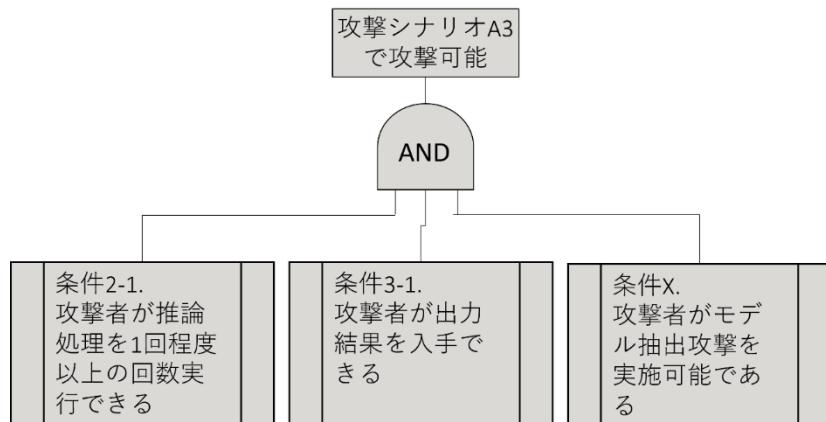


図 II- 11. 回避攻撃（敵対的サンプル）の攻撃シナリオ A3 のアタックツリーと攻撃実施可能条件

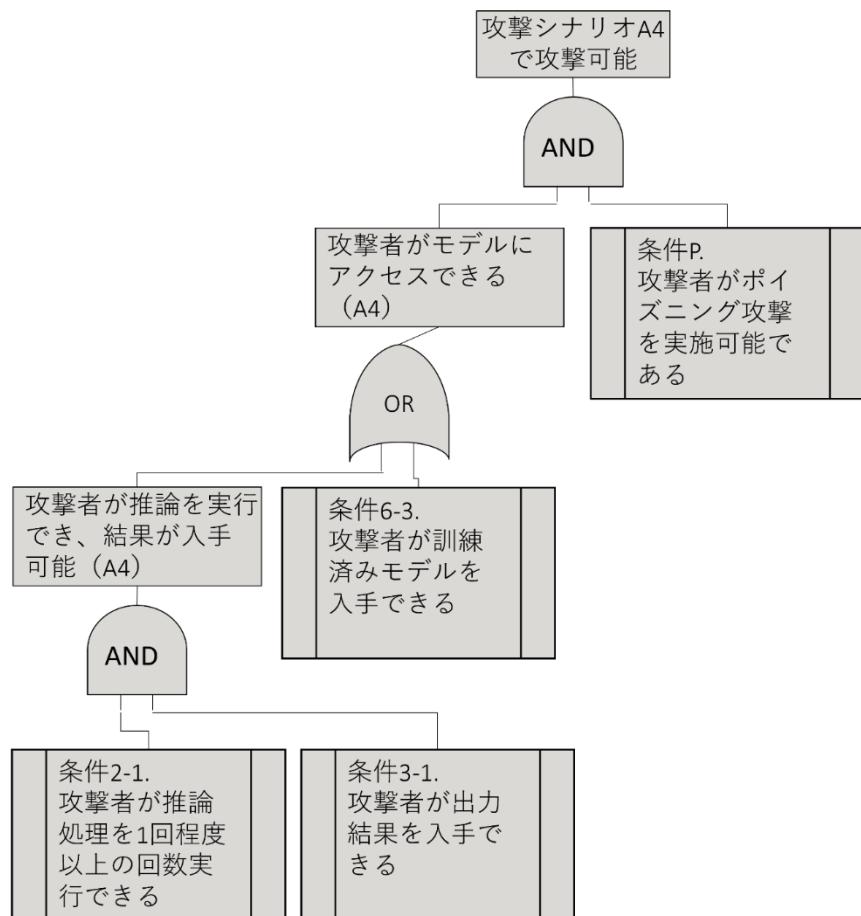


図 II- 12. 回避攻撃（敵対的サンプル）の攻撃シナリオ A4 のアタックツリーと攻撃実施可能条件

### 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。

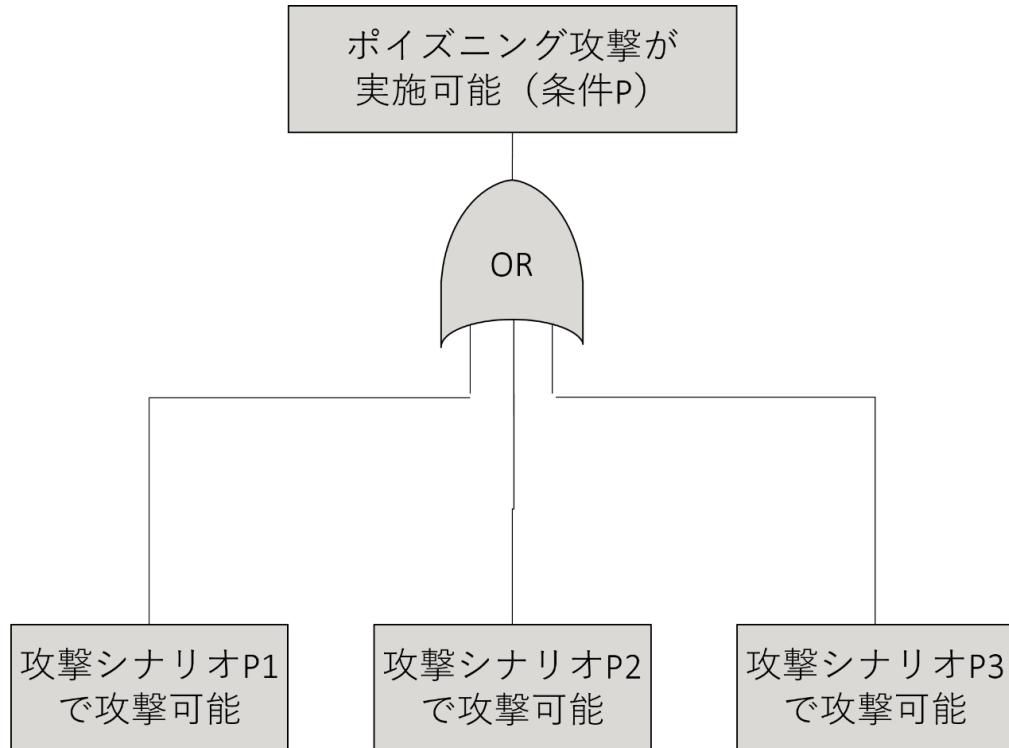


図 II- 13. ポイズニング攻撃のアタックツリー（上位部分）

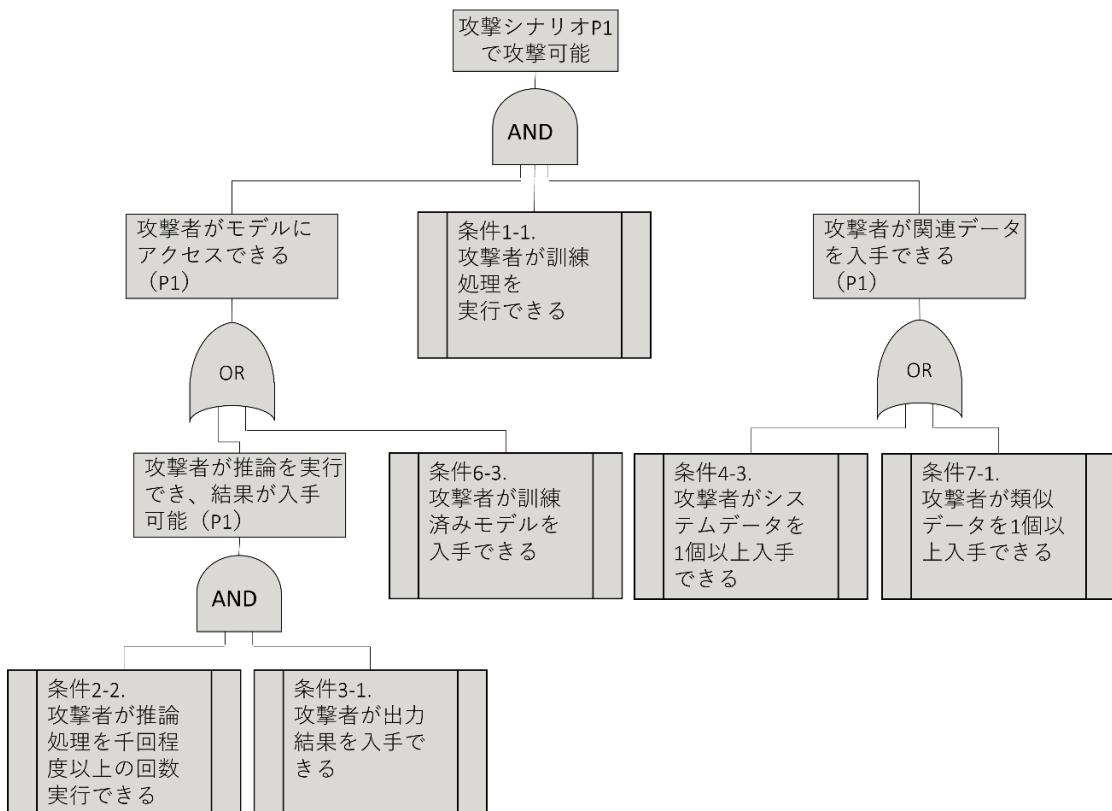


図 II- 14. ポイズニング攻撃の攻撃シナリオ P1 のアタックツリーと攻撃実施可能条件

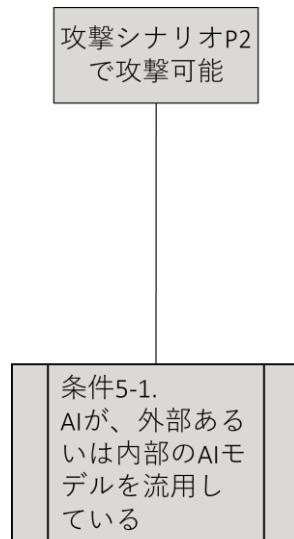


図 II- 15. ポイズニング攻撃の攻撃シナリオ P2 のアタックツリーと攻撃実施可能条件

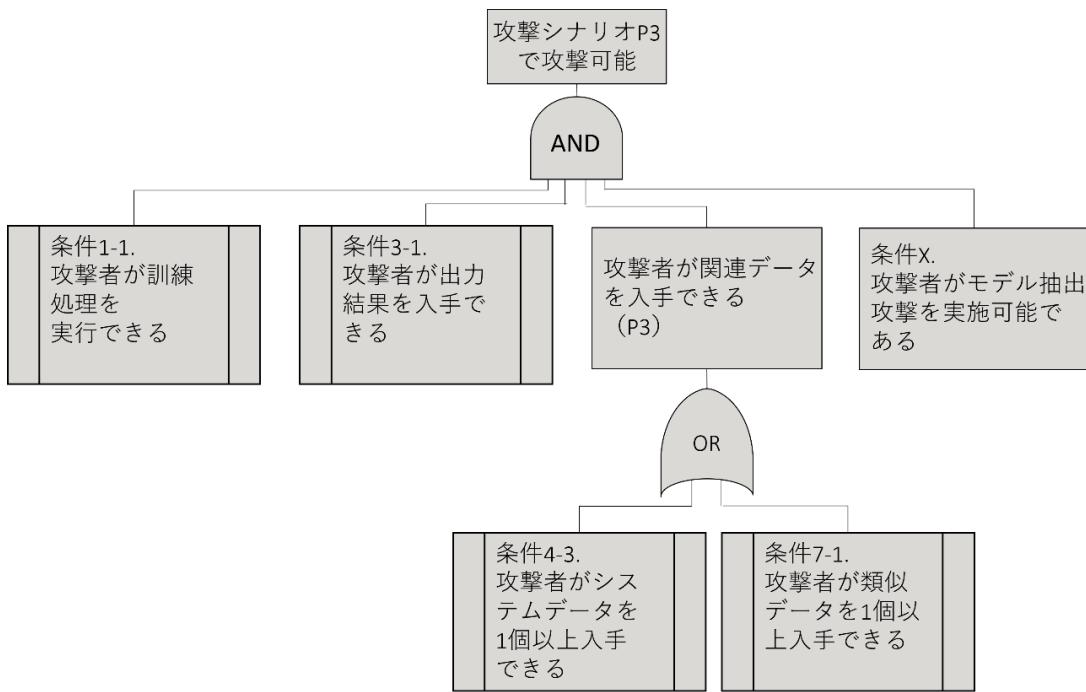


図 II- 16. ポイズニング攻撃の攻撃シナリオ P3 のアタックツリーと攻撃実施可能条件

### 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。

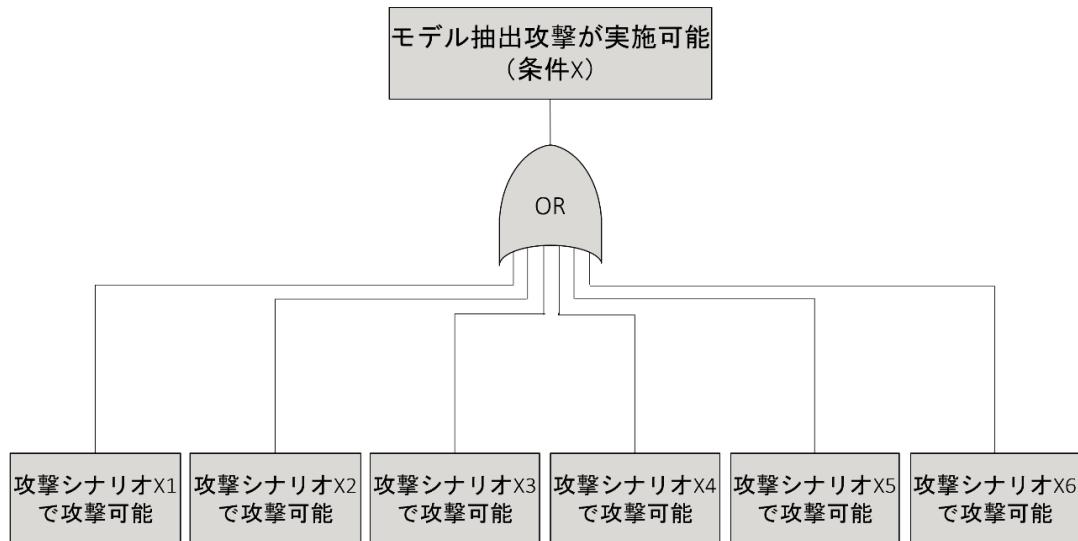


図 II- 17. モデル抽出攻撃のアタックツリー（上位部分）

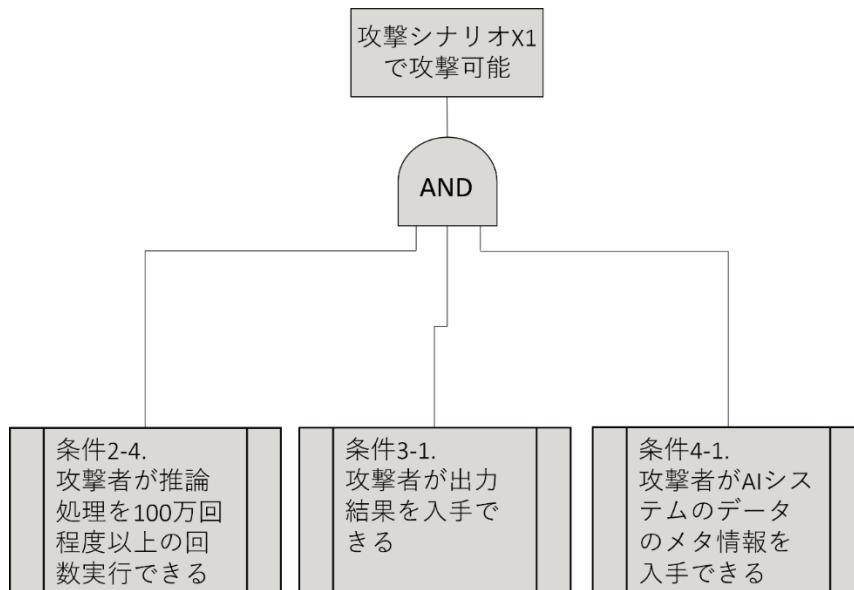


図 II- 18. モデル抽出攻撃の攻撃シナリオ X1 のアタックツリーと攻撃実施可能条件

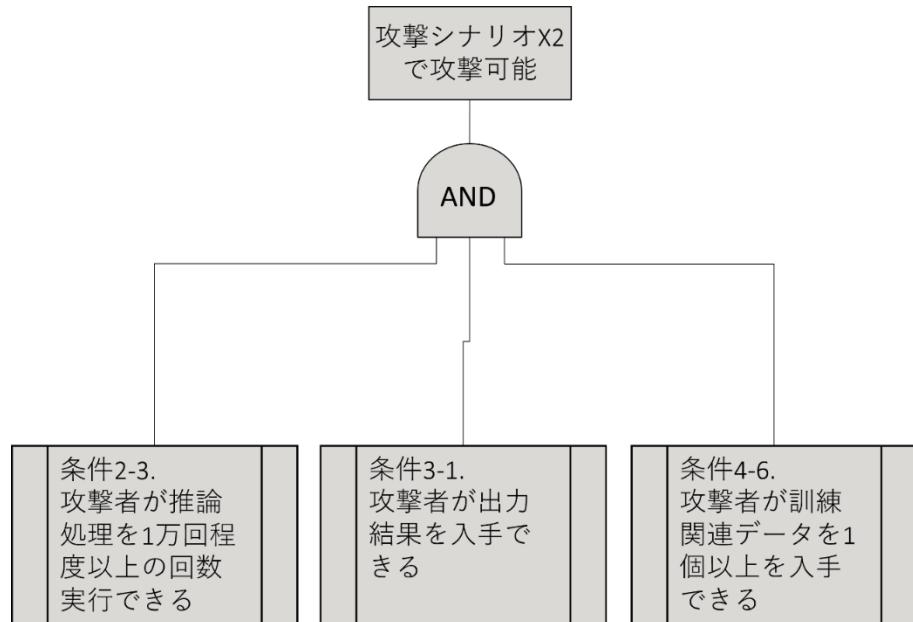


図 II- 19. モデル抽出攻撃の攻撃シナリオ X2 のアタックツリーと攻撃実施可能条件

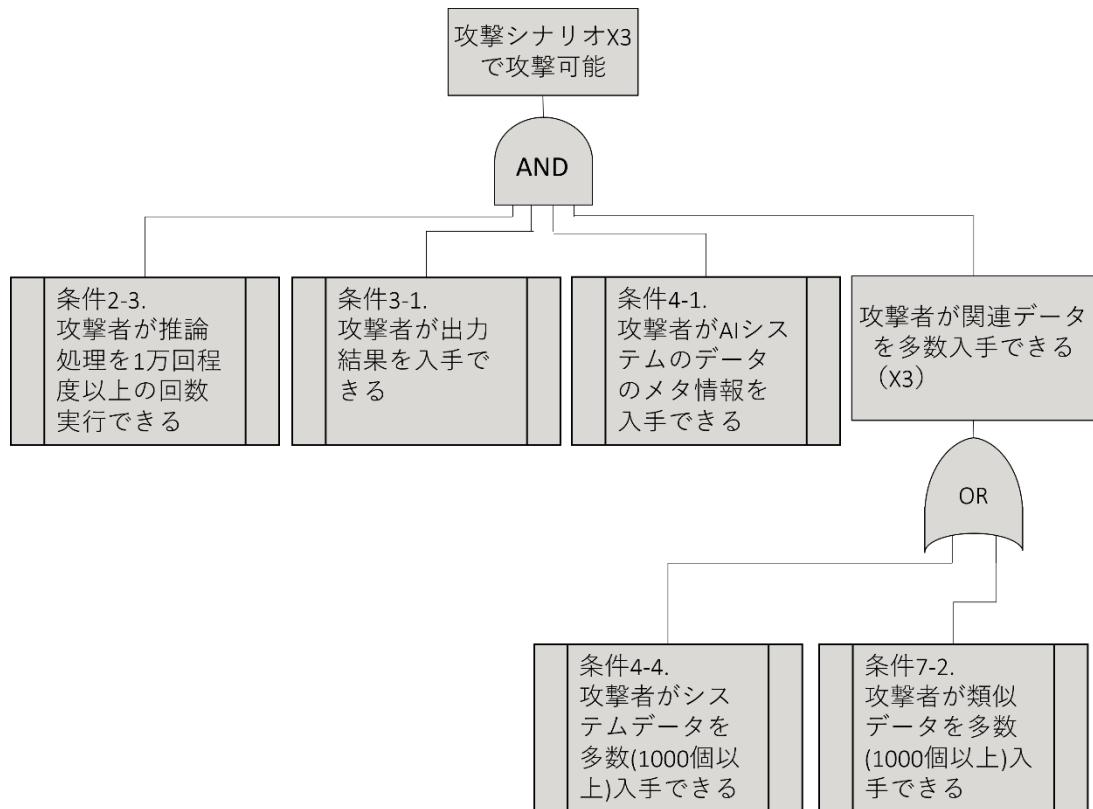


図 II- 20. モデル抽出攻撃の攻撃シナリオ X3 のアタックツリーと攻撃実施可能条件

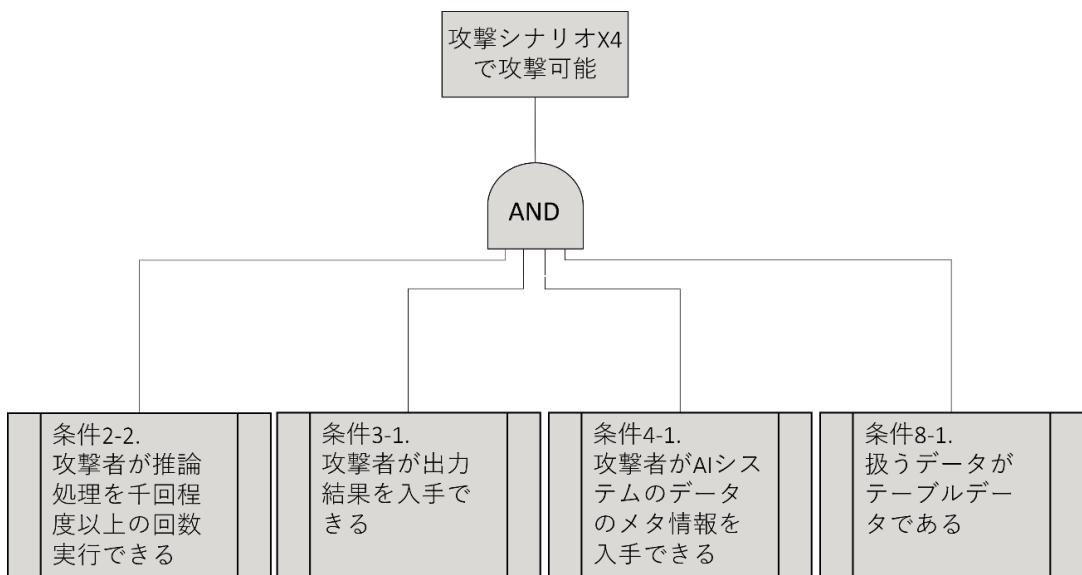


図 II- 21. モデル抽出攻撃の攻撃シナリオ X4 のアタックツリーと攻撃実施可能条件

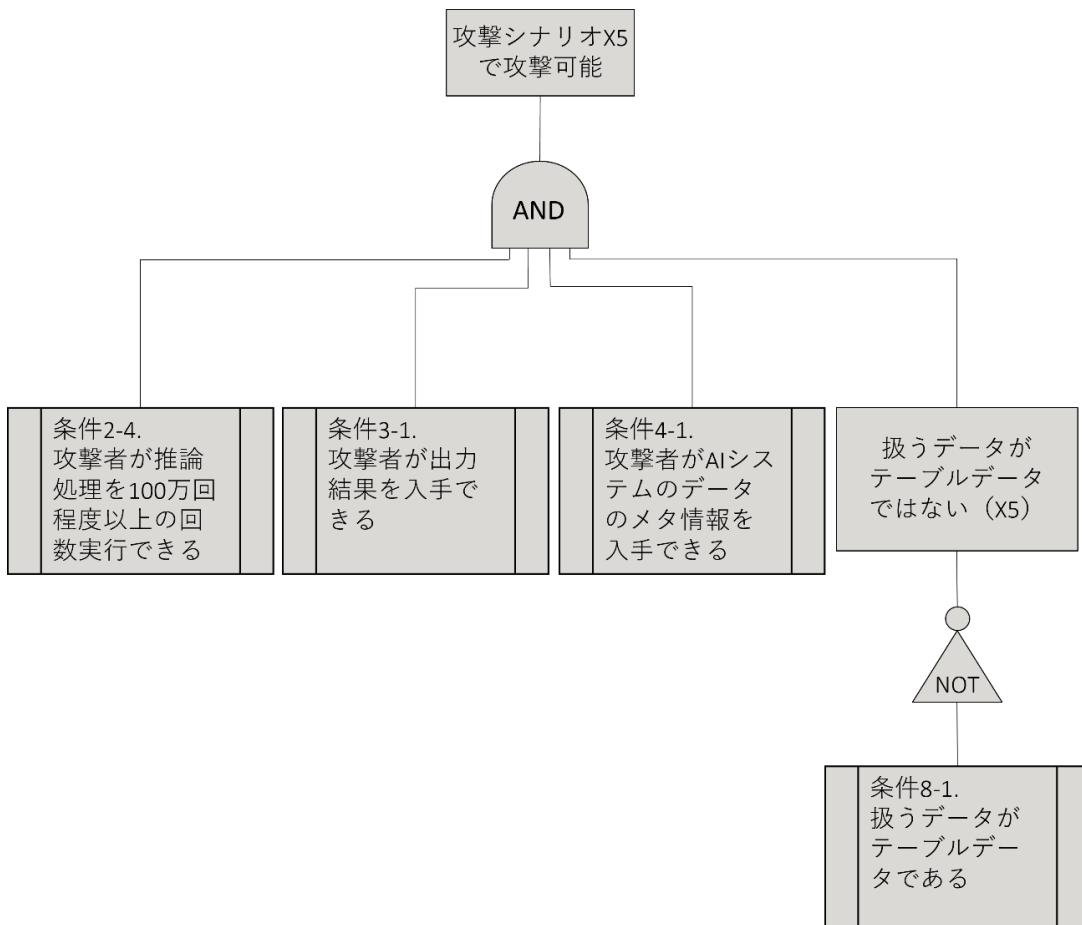


図 II- 22. モデル抽出攻撃の攻撃シナリオ X5 のアタックツリーと攻撃実施可能条件



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。

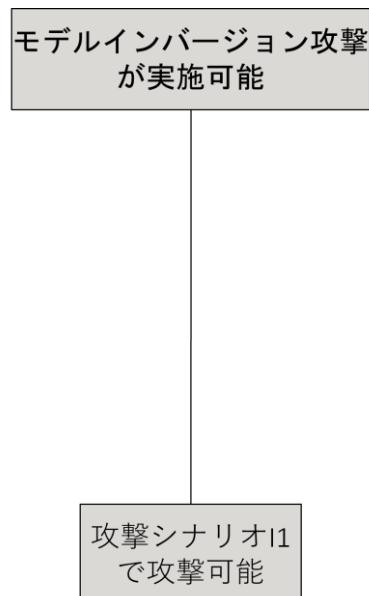


図 II- 24. モデルインバージョン攻撃のアタックツリー（上位部分）

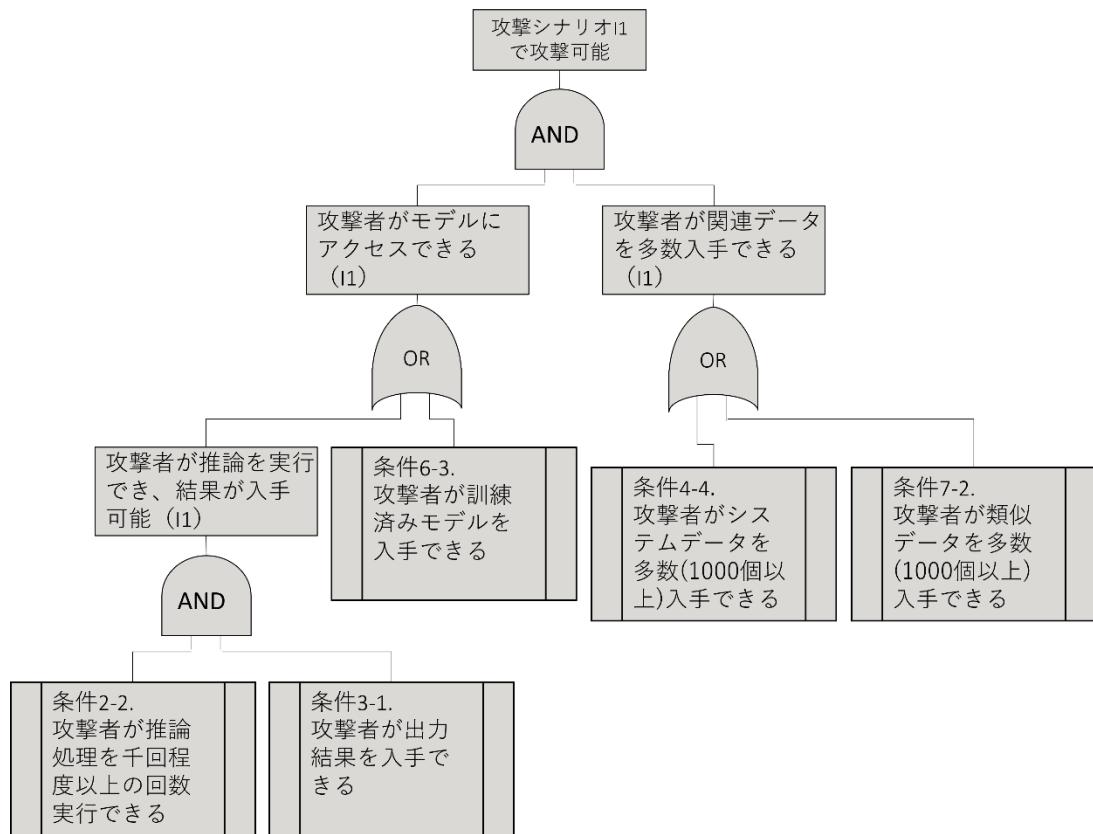


図 II- 25. モデルインバージョン攻撃の攻撃シナリオ I1 のアタックツリーと  
攻撃実施可能条件

### 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。

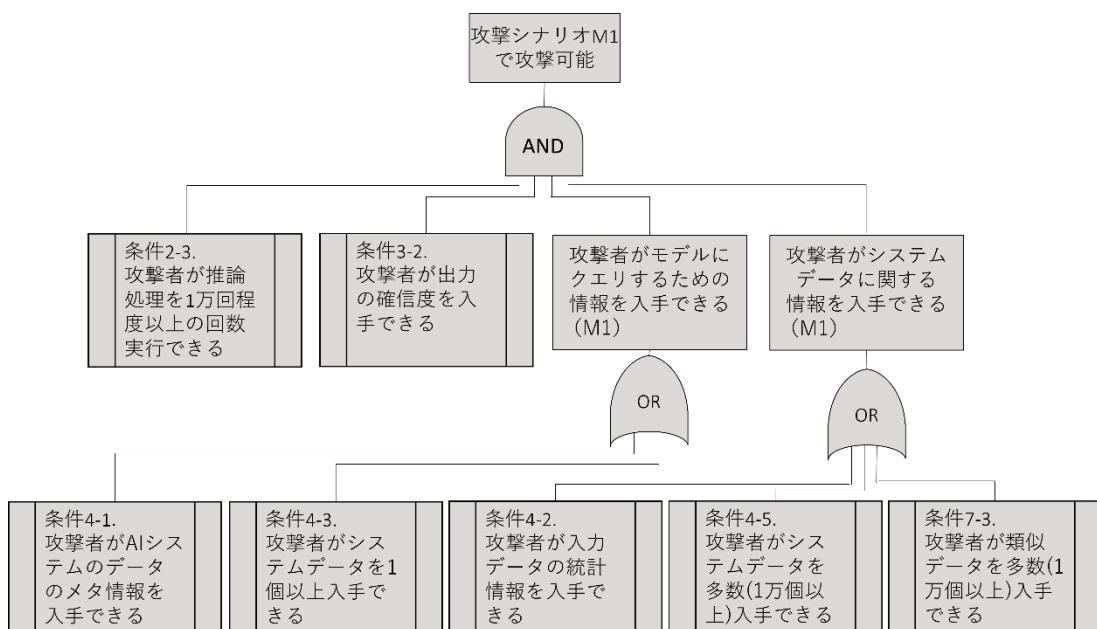
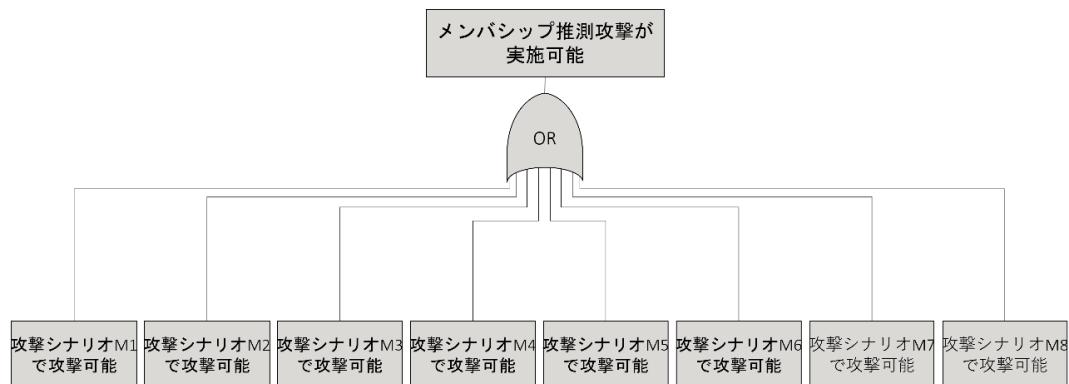


図 II- 27. メンバシップ推測攻撃の攻撃シナリオ M1 のアタックツリーと攻撃実施可能条件

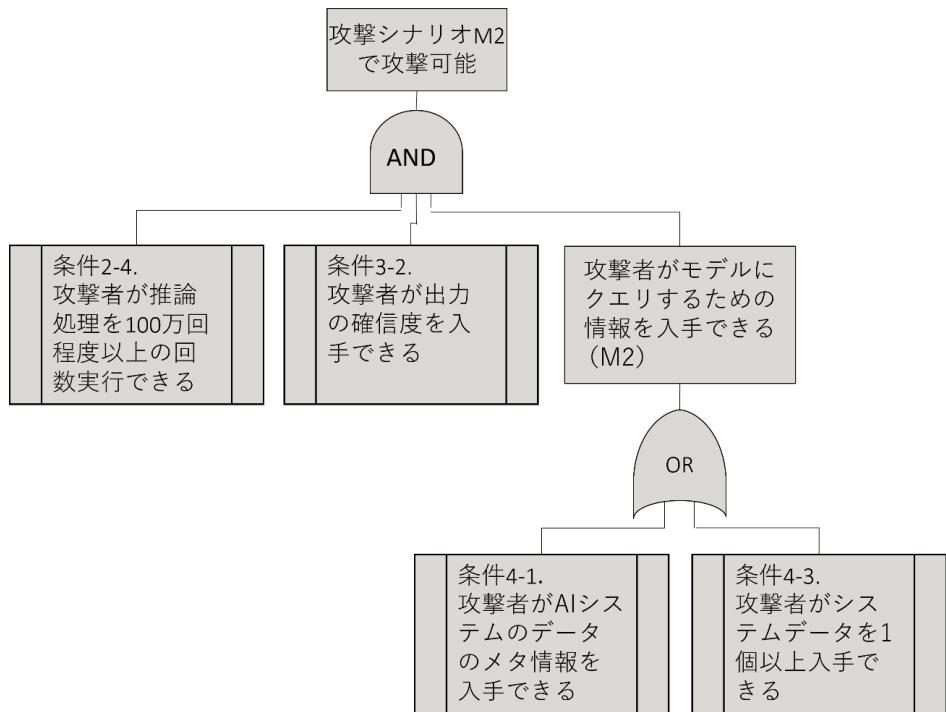


図 II- 28. メンバシップ推測攻撃の攻撃シナリオ M2 のアタックツリーと攻撃実施可能条件

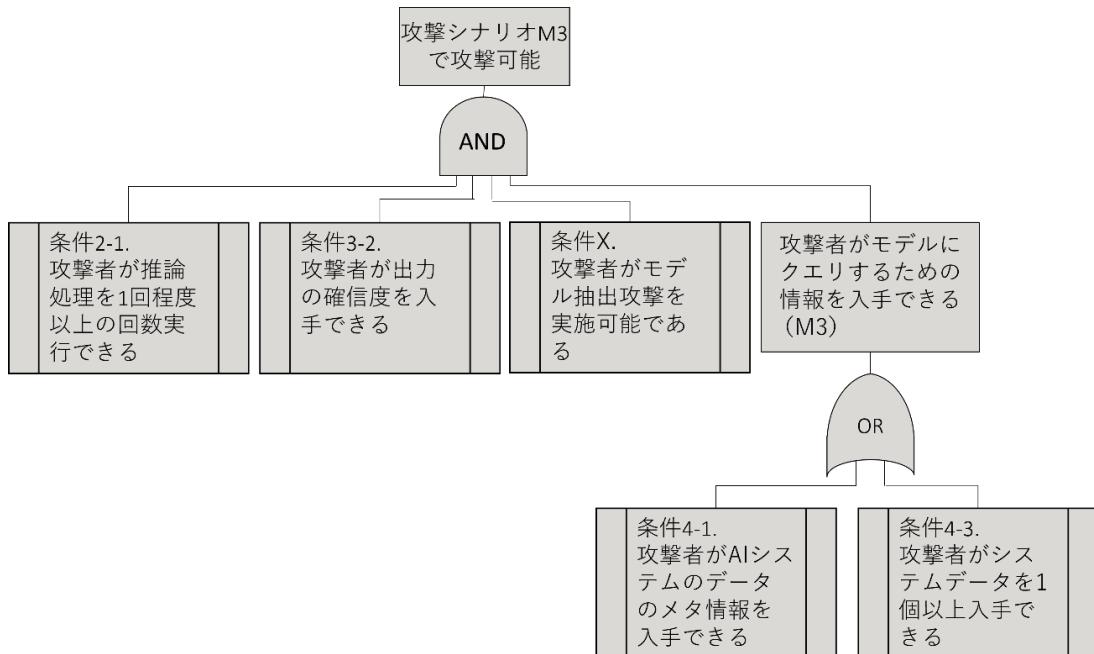


図 II- 29. メンバシップ推測攻撃の攻撃シナリオ M3 のアタックツリーと攻撃実施可能条件

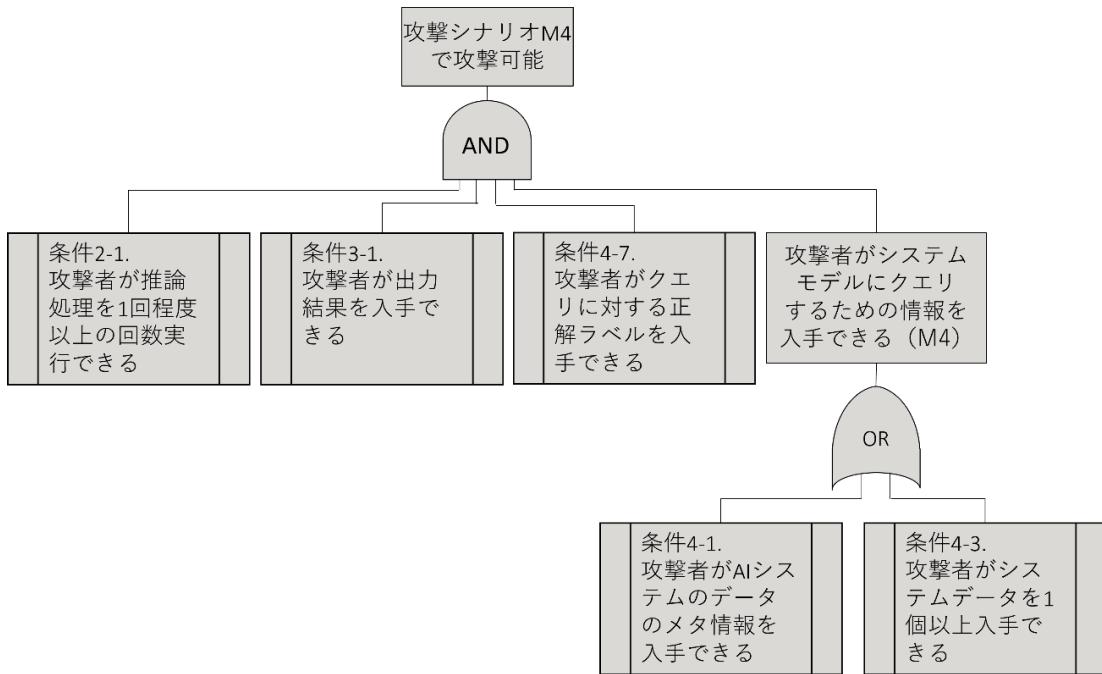


図 II- 30. メンバシップ推測攻撃の攻撃シナリオ M4 のアタックツリーと  
攻撃実施可能条件

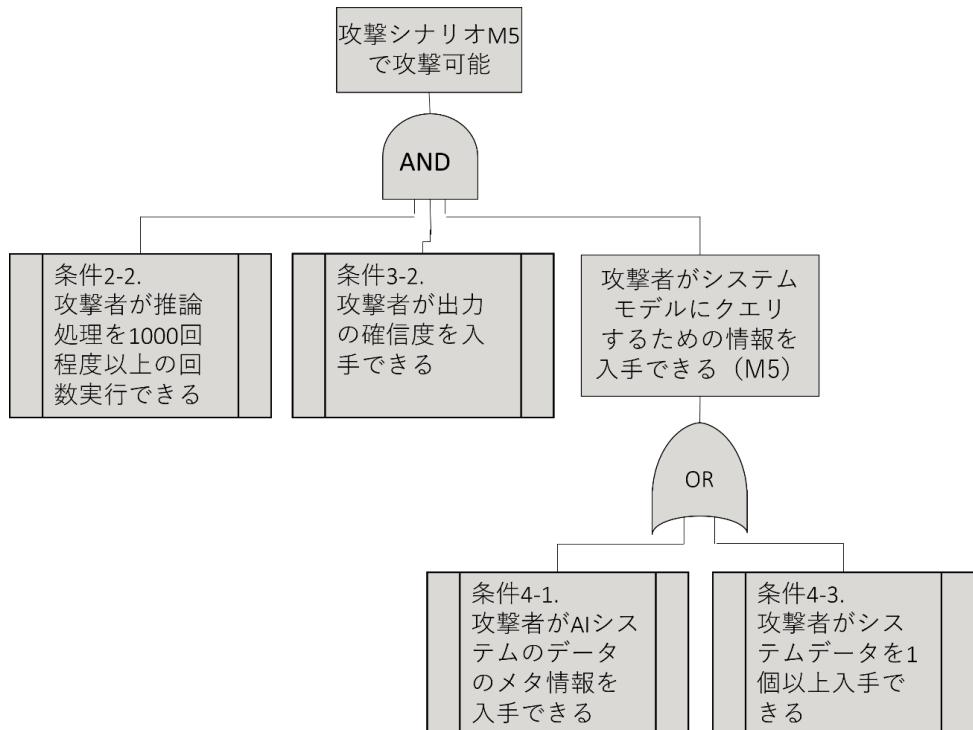


図 II- 31. メンバシップ推測攻撃の攻撃シナリオ M5 のアタックツリーと  
攻撃実施可能条件

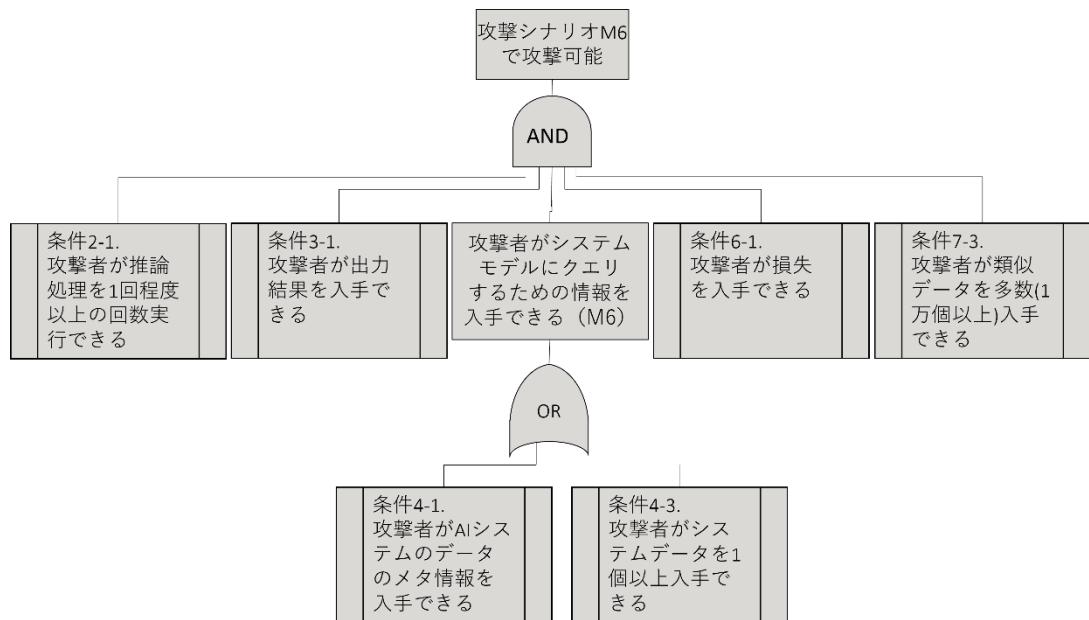


図 II- 32. メンバシップ推測攻撃の攻撃シナリオ M6 のアタックツリーと  
攻撃実施可能条件

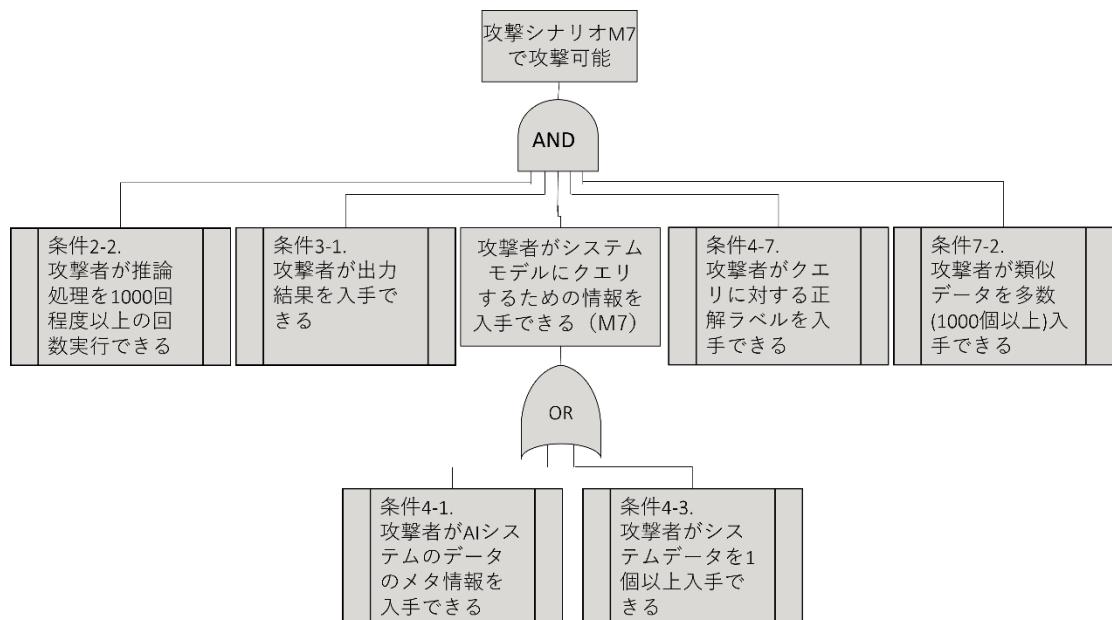


図 II- 33. メンバシップ推測攻撃の攻撃シナリオ M7 のアタックツリーと  
攻撃実施可能条件

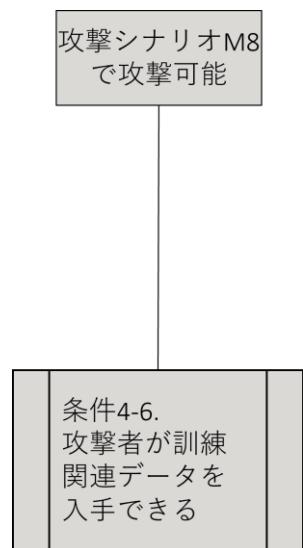


図 II- 34. メンバシップ推測攻撃の攻撃シナリオ M8 のアタックツリーと  
攻撃実施可能条件

### 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。テープルデータなら 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 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 7－3. 【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 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	類似データを千個以上入手可能	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 1 万個以上入手可能	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 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ます。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 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？	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？	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？	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
7-3	【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 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	【7-2】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 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 千個以上入手 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 1 万個以上入 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 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	Yes

	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？	-
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
7-3	【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 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を入手可能	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	類似データを千個以上入手可能	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 1 万個以上入手可能	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 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、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？	No
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 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 1,000 個程度以上、	No

	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 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	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	類似データを千個以上入手可能	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1万個以上入手可能	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 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身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
2-2B	【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	Yes

	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 1,000 個程度以上、何らかの手段で準備・入手できますか？	No
7-3	【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 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 千個以上入手 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 1 万個以上入 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 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 1,000 個程度以上、何らかの手段で準備・入手できますか？	Yes
7-3	【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 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 4-4	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 4-5	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	TRUE	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	TRUE	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	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 千個以上入手可能	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 1 万個以上入手可能	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 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	-

	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？	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 1,000 個程度以上、何らかの手段で準備・入手できますか？	Yes
7-3	【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 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	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	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 千個以上入手可能	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 1 万個以上入手可能	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 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 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？	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	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？	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見たがる事ができますか？	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見たがる事ができますか？	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 1,000 個程度以上、何らかの手段で準備・入手できますか？	Yes
7-3	【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 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	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 千個以上入手 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 1 万個以上入 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 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：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：M1, 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？	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	No

	(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？	
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 1,000 個程度以上、何らかの手段で準備・入手できますか？	Yes
7-3	【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 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	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 4-4	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 4-5	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	類似データを千個以上入手可能	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1万個以上入手可能	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 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](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策定委員会メンバーリスト

・現メンバー

乾 真季 (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)  
久連石 圭 (株式会社東芝)  
辻 健太郎 (富士通株式会社)  
花谷 嘉一 (株式会社東芝)  
林 昌純 (帝京平成大学)  
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)

・元メンバー

市原 大暉 (株式会社 NTT データ)  
及川 孝徳 (富士通株式会社)  
金子 朋子 (国立情報学研究所)  
田口 研治 (国立情報学研究所)  
森川 郁也 (富士通株式会社)  
(敬称略・五十音順)