機械学習システム
セキュリティガイドライン
「本編」

Version 0.8

2022 年 4 月 4 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・セーフ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

1. ガイドライン概要

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、2 編のガイドライン「本編」「リスク分析編」と 1 つの調査報告「攻撃検知技術の概要」で構成される（図 1）。

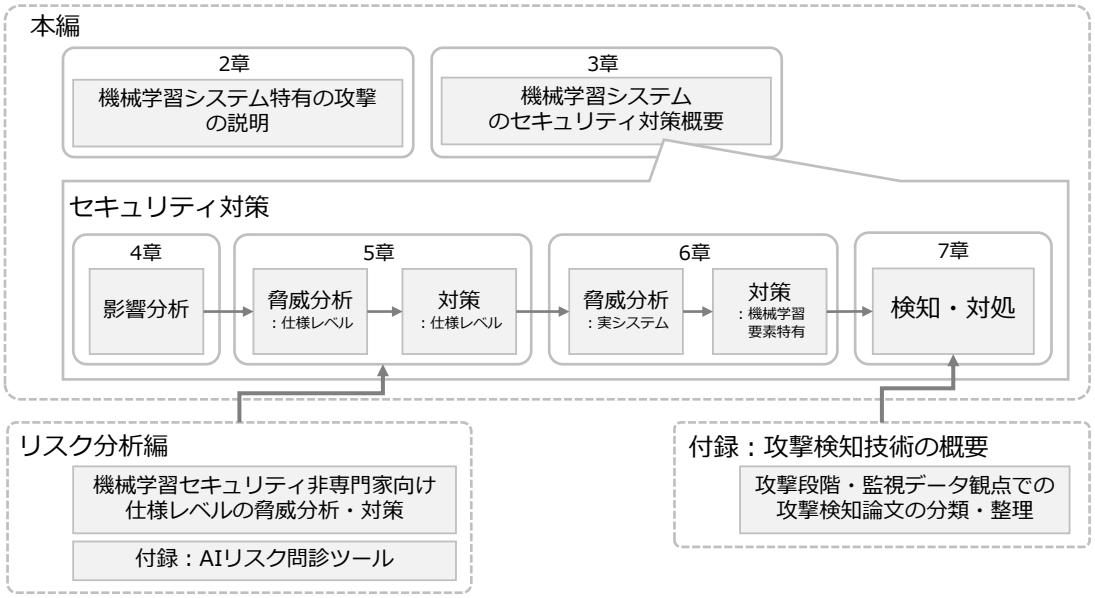

図 1　ガイドライン概要

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。2 章で機械学習システム特有の攻撃について説明し、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。4〜7 章では 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、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、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。

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[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」[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」[3]では図 2 のように整理している。

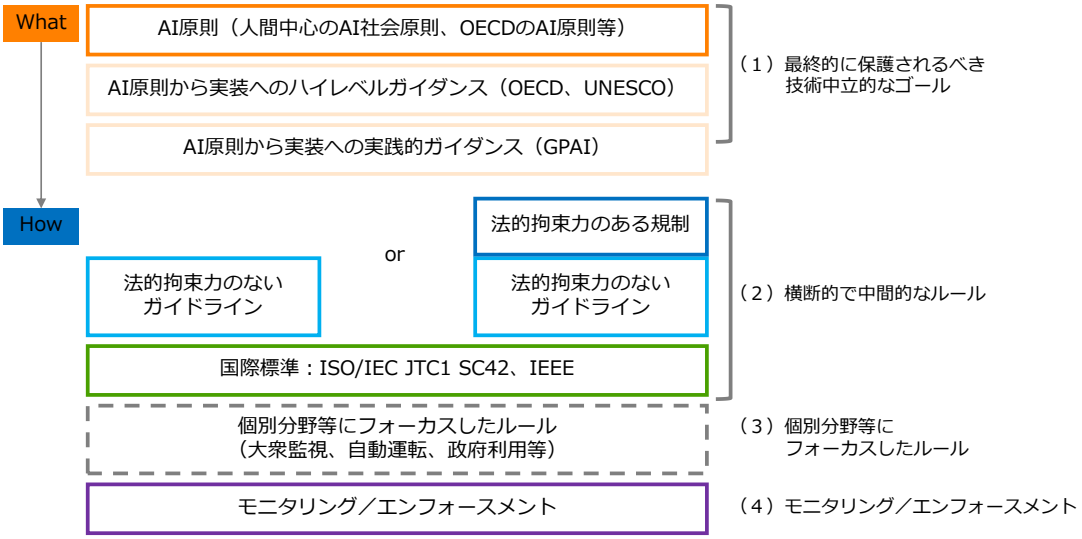

図 2　AI ガバナンスの構造

本ガイドラインは図 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。

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。

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を切り分けて説明する。

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

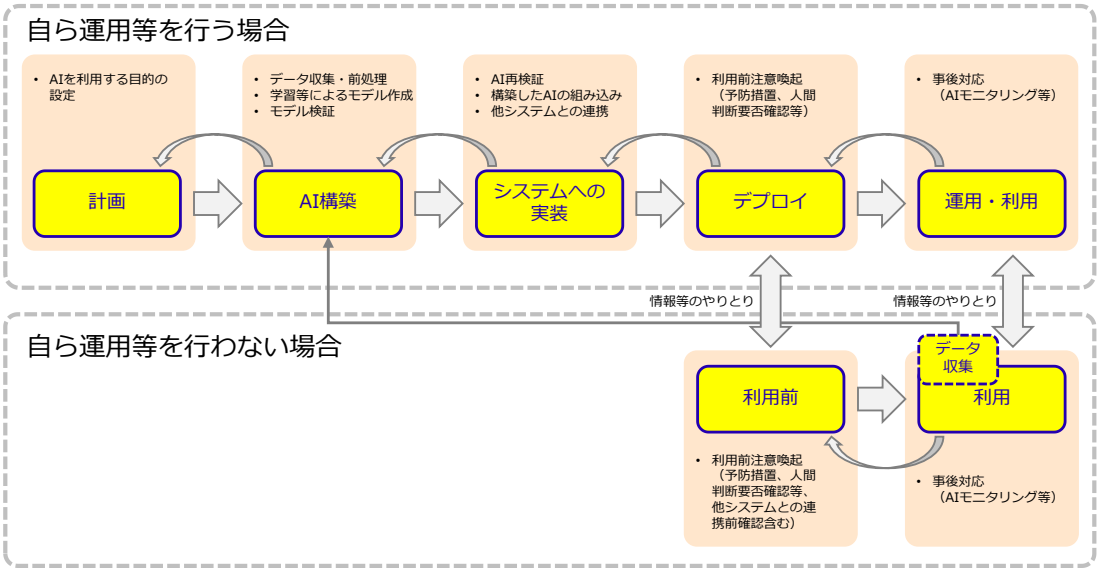

図 3　AI の利活用の流れ

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。

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[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 [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[6]や ICO の AI and data protection risk mitigation and management toolkit[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[9]、産総研[10]、ENISA[6]、ICO[8]等の機関や Microsoft[6]、三井物産セキュアディレクション株式会社[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。

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。

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 4・図 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。

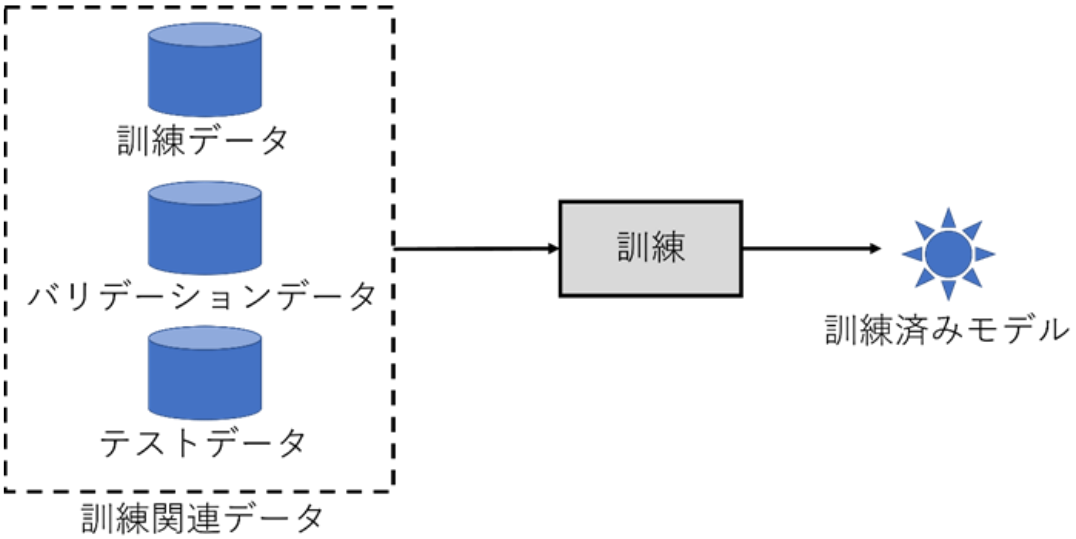

図 4 機械学習処理部の訓練パイプライン

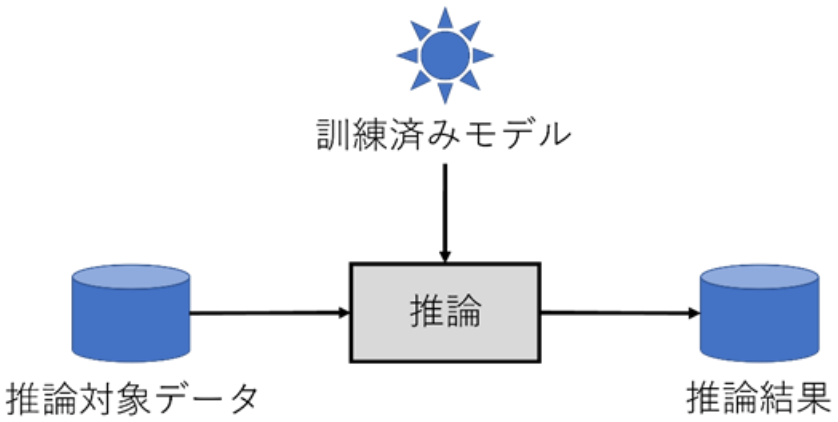

図 5 機械学習処理部の推論パイプライン

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[9]や産総研の機械学習品質マネジメントガイドライン[10]、Microsoft の公開している Threat Modeling AI/ML Systems and Dependencies[6]においても整理されており、参考となる。

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

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（2.2.1 項）や訓練データ・訓練モデルを汚染する攻撃（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。

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（2.2.3 項）がある。

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(2.2.4 項、2.2.5 項）がある。

2.2. 脅威を引き起こす攻撃

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。

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（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。

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（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。

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

本攻撃はモデルの窃取（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。

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（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。

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（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。

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、これまでも対応され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。

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うる脅威（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（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 6 に示す。

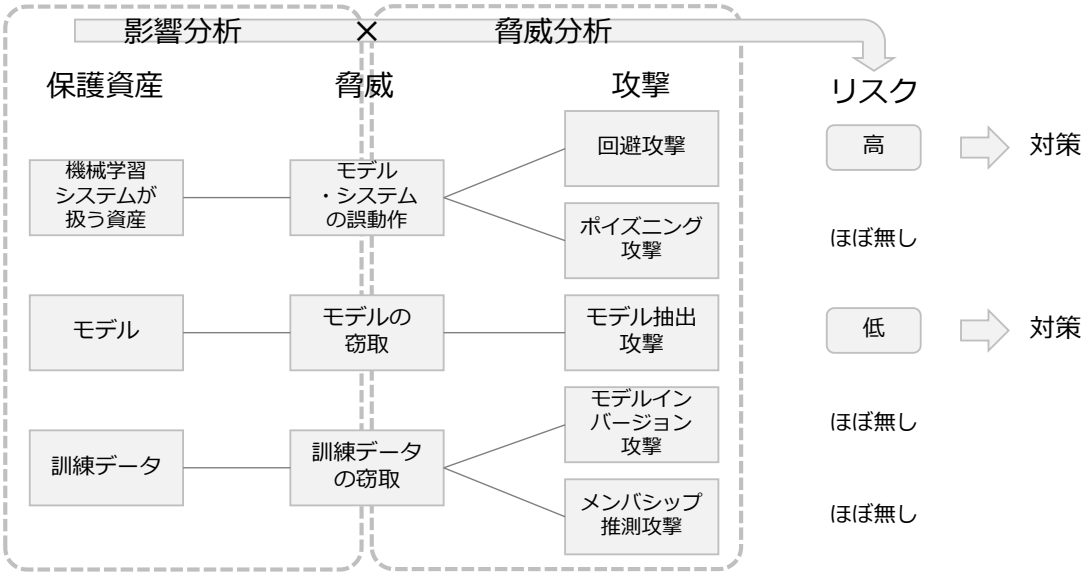

図 6 機械学習セキュリティの全体イメージ

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った未知の脅威への対策として、「検知・対処」を実施する。

3.2. 進め方

前節各工程の進め方の例を図 7 に示す。

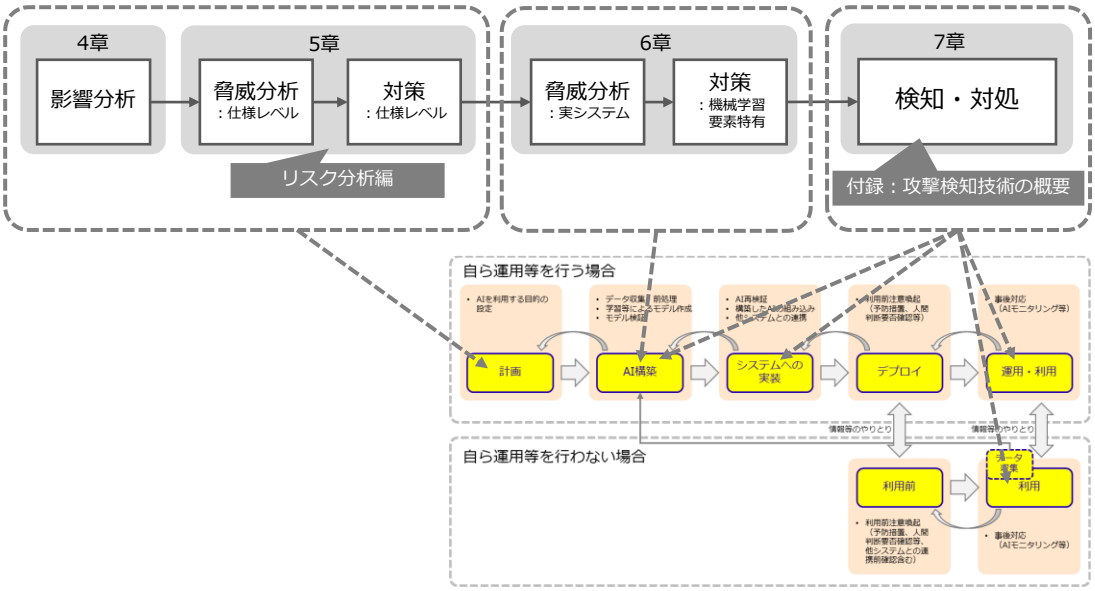

図 7 機械学習セキュリティの工程

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時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。

3.3. 各工程の実施について

前節の各工程の実施おいては、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。

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。

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 [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。

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

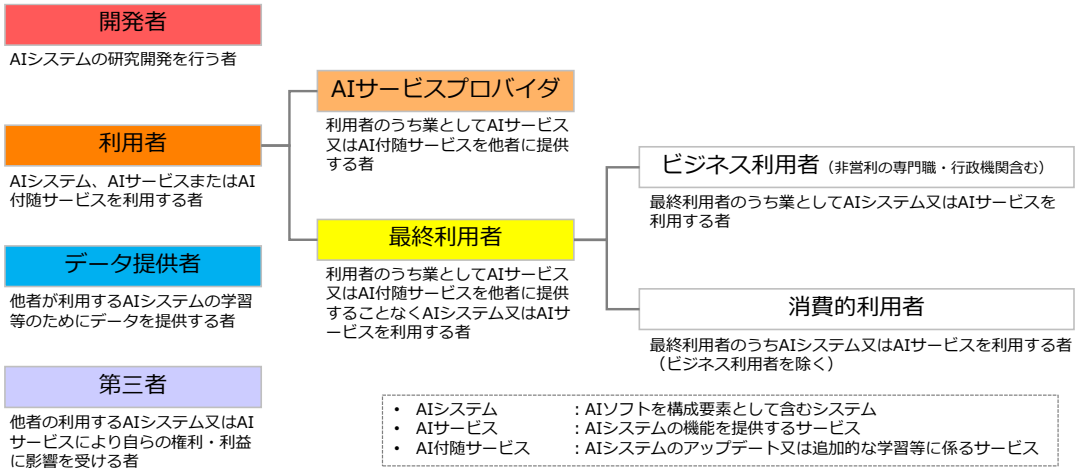

図 8 AI の利活用において関与が想定される主体

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、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 9 に示す。

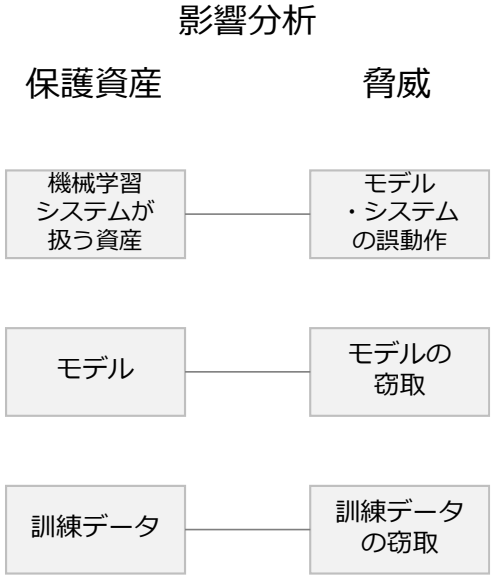

図 9 保護資産と脅威の結びつけ

「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[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。

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[4]の AI 利活用の流れにおける、「計画」フェーズ（図 3）で行うことを想定している。

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。

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（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係する主体（4.2 節）の中から想定攻撃者を設定し（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（5.1.2 項）する。各工程について説明する。

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係する主体（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係する主体の中で、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 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[4]では AI の利活用において関与が想定される主体を図 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。

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。

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・

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 1 に挙げる。

表 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	➢ 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	➢ 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 ✧ 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	➢ 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	➢ 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。

5.2. 仕様レベルでの対策

4 章の影響分析の結果と 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、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、そもそもの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 2 に各緩和策の例を示す。

表 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	➢ 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 ✧ 入力回数に制限をつける ➢ 必要以上の情報を出力しない ✧ 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	➢ 訓練に利用するデータ・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	➢ リスクを許容するシステム設計に変更する ✧ 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。

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[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を想定している。

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（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。

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（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（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)[12]、CleverHans[13]、Counterfit[14]などが公開されている。

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 3 に示す。緩和策実施後は、再度 6.1 節の脅威分析を行い、効果を評価・確認する。

表 3　機械学習要素特有の緩和策例

攻撃	緩和策例
回避攻撃	➢ 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摂動のサイズを決め、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	➢ 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	➢ 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	➢ 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	➢ 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

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[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、５，６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。

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[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。

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[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[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。

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 4 に挙げる。

表 4　攻撃毎の攻撃事象例

攻撃	前兆検知		兆候検知
回避攻撃	➢ 対象システムの調査 ✧ 分類ラベル数 ✧ クエリ上限 ➢ 攻撃用アカウントの作成	➢ 敵対的サンプルを作成する活動	➢ 敵対的サンプルの入力
ポイズニング攻撃		➢ 汚染データの訓練データへの混入 ➢ 汚染モデルの混入	➢ トリガーを含む入力（バックドア攻撃）
モデル抽出攻撃		-	➢ 攻撃クエリの入力
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。

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、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。

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 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 5 応急対策例

分類	例
システムの制限・停止	➢ 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 （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） ✧ 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	➢ ノイズ除去等による摂動緩和（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	➢ 出力ラベルの偽装 （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 （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）

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 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	➢ 攻撃対象ラベルの特定 ➢ 敵対的サンプルの作成手法特定	➢ 過去の回避攻撃探索	➢ 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	➢ 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	➢ 過去のポイズニング攻撃が原因の誤判定探索	➢ 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	➢ 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	➢ 過去の攻撃試行の探索	-
モデルインバージョン攻撃	➢ 漏えいした訓練データの特定 ➢ 攻撃手法の特定	➢ 過去の攻撃試行の探索	➢ 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	➢ 漏えいした情報の特定 ➢ 攻撃手法の特定	➢ 過去の攻撃試行の探索	➢ 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

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

8. 参考文献

[1] Organisation for Economic Co-operation and Development (OECD),
Principles on Artificial Intelligence.
https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[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

[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

[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

[5] MITRE, ATLAS. https://atlas.mitre.org

[6] European Network and Information Security Agency (ENISA),
Artificial Intelligence Cybersecurity Challenges.
https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

[7] Microsoft Corporation, Threat Modeling AI/ML Systems and Dependencies.
https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml

[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/

[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

[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

[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

[12] International Business Machines Corporation, Adversarial Robustness Toolbox.
https://github.com/Trusted-AI/adversarial-robustness-toolbox

[13] CleverHans Lab, CleverHans.
https://github.com/cleverhans-lab/cleverhans

[14] Microsoft Corporation, Counterfit.
https://github.com/Azure/counterfit/

[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
[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）