ICSCoE TLP: WHITE

セキュリティ関係者のための AIハンドブック

2022 年 6 月

独立行政法人 情報処理推進機構

産業サイバーセキュリティセンター

中核人材育成プログラム 5 期生

チーム Security & AI

目次

はじめに

近年、機械学習（Machine Learning; ML）をはじめとした人工知能（Artificial Intelligence; AI） 技術の進歩は著しい。AI 技術は、重要インフラを含む様々な産業・事業領域において、デジタルトランスフォーメーション（DX）の推進などに活用されている。

その一方で、デジタル化の進展とともにサイバー攻撃のリスクが日に日に高まっている。あらゆる事業にデジタルデータが用いられる現在、サイバー攻撃は組織にとって事業継続を脅かす重大なリスクである。

AI 技術の発展とサイバーセキュリティリスクの増大は互いに無関係ではない。AI を活用したセキュリティ製品も普及してきている。一方、サイバー攻撃にさらされる AI システムは増加してきている。
セキュリティ関係者にとって、AI・機械学習を知ることは、組織のセキュリティを向上させる上で重要である。

例えば、機械学習では、一般的にブラックボックスで処理が行われ、また学習用データにより精度が変化するなどの特徴がある。そのため、AI を活用したセキュリティ製品が誤検知を起こした際に、その誤検知に至った理由が分からず、説明できないという事態が生じうる。
また、AI システムは、学習用データを収集する、AI に学習用データを学習させるというプロセスを必要とする。そのため、AI システムの開発～運用プロセスは通常のシステム開発とは異なる。

こうした AI 固有の特徴を理解してセキュリティ対策の導入を進めるとともに、AI 固有の特徴を悪用したサイバー攻撃が起こりうることも認識する必要がある。そのため、セキュリティ関係者には今後より一層 AI に関する知識が求められるようになる。
しかし、セキュリティと AI は共に専門性の高い分野であり、セキュリティ関係者が読めるような、両分野を包括的に説明する資料は国内ではあまり見かけない。

そこで、各組織でのセキュリティ分野における適切な AI の活用と、組織で活用される AI のセキュリティ向上を推進することを目的として、セキュリティ分野における AI の利用とリスクに関する情報を整理する本書を執筆した。

各組織のセキュリティ関係者に本書をお読みいただき、少しでも目的が達成されれば幸いである。

第1章　　本書について

1.1　　目的

本書は、セキュリティ分野における AI の利用とリスクに関する情報を整理することで、各組織でのセキュリティ分野における適切な AI の活用と、組織で活用される AI のセキュリティ向上を推進するための参考資料としていただくことを目的とする。

1.2　　「セキュリティと AI」3 つの領域

内閣サイバーセキュリティセンター（NISC）や、欧州サイバーセキュリティ機関（ENISA）が発行する文書［1] [2]では、セキュリティと AI の関係性を次の 3 つに分類している[1]。本書もこの分類を採用する。

- ✓ AI を活用したサイバーセキュリティ対策…（第 3 章で説明）
- ✓ AI そのものを守るセキュリティ…（第 4 章で説明）
- ✓ AI を使ったサイバー攻撃…（コラムで説明）

1.3　　本書における「AI」という言葉の使い方

「AI」とは Artificial Intelligence の略称であり「人工知能」と訳される。近年では AI の一部である「機械学習（Machine Learning; ML）」、および機械学習の一部である「深層学習（Deep Learning; DL）」が注目を浴びており、AI という用語も機械学習と同義で使われている場面も多い［1]。

本書は、セキュリティと AI の観点で、AI の中でも近年特に使用例が多い「機械学習」「深層学習」に焦点を当てた文書である。そのため、本書において「AI」という用語は、個別に説明する場合を除き、機械学習および深層学習を指すものとする。ただし、特に機械学習や深層学習であることを強調する場合には、「機械学習」「深層学習」「AI（機械学習）」のように表記する。

また、2.3 節で説明するが**本書においては「AI はブラックボックスである」という立場で議論を進めていく。**

「AI」「機械学習」「深層学習」とこれらの関係は第 2 章で説明する。

[1] 「AI 自身による攻撃」を含め 4 分類としている場合もある。例えば、総務省 サイバーセキュリティタスクフォース事務局「今後重点的に取り組むべき研究開発課題について」など <https://www.soumu.go.jp/main_content/000666230.pdf>

1.4　本書の構成

- ✓ 第 1 章では、本書の構成について説明する。
- ✓ 第 2 章では、第 3 章・第 4 章の内容を理解するために必要と考えられる、AI（機械学習）とそれに関連する事項について説明する。**読者に AI（機械学習）の知識がある場合、この章は読み飛ばしてかまわない。**
- ✓ 第 3 章では、「AI を活用したサイバーセキュリティ対策」を組織で企画・導入・運用する際の留意点について説明する。
- ✓ 第 4 章では、「AI そのものを守るセキュリティ」について説明する。
- ✓ 第 5 章では、セキュリティと AI に関連する事項について説明する。AI 社会原則や AI ガバナンス、AI 倫理、AI 品質、プライバシー、MLOps など、セキュリティとよく一緒に議論される内容について説明する。
- ✓ 第 6 章では、本書内容をまとめ、また、課題と制約について説明する。

1.5　対象とする読者

本書は、主に次の内容に興味や関心がある方を対象としている。

- 「AI を活用したサイバーセキュリティ対策」の企画・導入・運用
- AI システムや AI そのものを守るセキュリティ

例えば、以下のような方（セキュリティ関係者）を読者として想定している。

＜ユーザー企業担当者＞

- 「セキュリティ企画・運用」や「CSIRT」に従事するセキュリティ担当者
- 「AI システムの企画・開発・運用」に従事し、セキュリティ検討を行うシステム担当者

＜AI システムを開発・提供しているベンダー＞

- 営業担当者、AI システム開発者、AI アルゴリズム開発者

＜セキュリティベンダー＞

- 営業担当者

1.6　免責事項

- 本書は単に情報として提供され、内容は予告なしに変更される場合がある。
- 本書に誤りがないことの保証や、商品性または特定目的への適合性の黙示的な保証や条件を含め明示的または黙示的な保証や条件は一切ないものとする。
- 本書に記載の内容は、独立行政法人 情報処理推進機構および産業サイバーセキュリティセンターの意見を代表するものではなく、著者の見解に基づいている。
- 本書の利用によるトラブルに対し、本書著者ならびに監修者は一切の責任を負わないものとする。
- 本書の有効期限は、発行日から 2 年間とする。

1.7　参考とした資料

本書では、各内容を検討するにあたり、執筆時点における内容の網羅性や情報の正確性を可能な限り担保するために、公的機関や、セキュリティや AI を専門とする独立性・公共性の高い組織が発行する文献を参考文献として使用している。

以下は、特に参考とした文献である。

- **ENISA ”Artificial Intelligence Cybersecurity Challenges [2]”**
  （以降 ”AI Cybersecurity Challenge”）
  AI（機械学習）のライフサイクル、資産、脅威、およびそれらのマッピングについて説明しているドキュメント。本書第 4 章の執筆にあたり参考とした。

- **ENISA ”Securing Machine Learning Algorithms [3]”**
  （以降 ”Securing ML Algorithms”）
  特に AI（機械学習）のアルゴリズムおよびモデルに対する脅威（攻撃）、脆弱性、対策、およびそれらのマッピングについて説明しているドキュメント。本書第 4 章の執筆にあたり参考とした。

- **G20「AI 原則［4]」**
- **内閣府 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［5]」**
- **経済産業省 AI 原則の実践の在り方に関する検討会「AI 原則実践のためのガバナンス・ガイドライン ver. 1.1 [6]」**
  AI を社会で適切に利用・実装していくための原則や、その原則の具体的実践について説明しているドキュメント。「AI 活用セキュリティ製品は AI のセキュリティ分野への応用である」という観点でこれらのドキュメントを確認し、本書第 3 章の執筆にあたり参考とした。

- **英国 National Cyber Security Centre “Intelligent security tools [7]”**
  AI を活用したサイバーセキュリティ対策（含む製品の企画・導入・運用）を行う上でのポイントを説明している Web サイト。本書第 3 章の執筆にあたり参考とした。

第2章　　セキュリティのための AI（機械学習）知識

本章では第 3 章、第 4 章の理解のために必要と思われる AI の基礎事項を説明する。**AI について知識がある読者は、本章を読み飛ばしていただいてかまわない。**

AI 自身や関連する技術・動向は日進月歩であるから全体像を正確かつ詳細に説明することは難しい。本書ではそれらの理論的および技術的詳細は深追いしないので、詳細については適宜、脚注や参考文献を参照していただきたい。

また、本書で用いている用語の使い方についても、世の中における明確な定義がない、もしくは使い方に幅がある場合がある。将来、使い方が変わる可能性も考えられる。本書の用語の使い方が全てではない点をご了承いただきたい。

2.1　AI とは

「**AI**」とは Artificial Intelligence の略称であり、「人工知能」と訳される。人工知能の明確な定義は存在しないが、一般社団法人 人工知能学会「設立趣意書」からの抜粋では「大量の知識データに対して、高度な推論を的確に行うことを目指したもの[2] [3]」、総務省「令和元年版 情報通信白書 [1]」によれば「人間の思考プロセスと同じような形で動作するプログラム全般」などとされている。

1.3 節でも述べたように、近年では AI は「機械学習」や機械学習の手法の一つである「深層学習」と同義で使われる場合も多い。本書もその用法にのっとることとしている。

ただし、そうした状況であっても、「機械学習」以外の AI も依然活用されている。例えば、製造業における生産管理スケジューリングなどに使われる[4]「遺伝的アルゴリズム[5]」やソースコード診断における「SMT ソルバー」[6]などは、「機械学習」以外の AI の活用事例である。

AI の活用の動向について、情報処理推進機構「DX 白書 2021 [8]」によれば、AI 技術を「導入している」と回答した日本企業は 20.5%にのぼる。「AI 白書 2020 [9]」の調査結果である 4.2%から大きく増加しており、AI の活用が広がってきていることがわかる。

また、「AI 白書 2022 [10]」によれば、AI 技術の便宜的な整理・分類として「知覚・認識・理解・意図」「学習」「身体性」「認知発達」「意識」「言語」「知識」「判断」「創作」の 9 つが挙げられており、技術や適用範囲の広さがうかがえる。

---

[2] 一般社団法人 人工知能学会「設立趣意書」<https://www.ai-gakkai.or.jp/about/about-us/jsai_teikan/>

[3] 厚生労働省資料 <https://www.mhlw.go.jp/file/05-Shingikai-10601000-Daijinkanboukouseikagakuka-Kouseikagakuka/0000148673.pdf>

[4] ビジネス＋IT 記事<https://www.sbbit.jp/article/cont1/36383>

[5] 生物進化を模した近似解探索手法。電気学会<https://www.iee.jp/pes/termb_093/>の解説など。

[6] ブロードバンドセキュリティ<https://www.sqat.jp/information/4117/>など。

2.2　機械学習、深層学習

本節ではまず、「機械学習」について説明する。AI 同様に、機械学習にも統一された明確な定義は存在しないようである。下に国内外の機関による機械学習の定義を列挙する。

「人間の学習に相当する仕組みをコンピューター等で実現するものであり、一定の計算方法（アルゴリズム）に基づき、入力されたデータからコンピューターがパターンやルールを発見し、そのパターンやルールを新たなデータに当てはめることで、その新たなデータに関する識別や予測等を可能とする手法」 **出典：総務省 「令和元年版 情報通信白書** [1]**」, 2019 年**

Machine learning (ML), which can be defined as the ability for machines to learn from data to solve a task without being explicitly programmed to do so,（略）
（機械学習（ML）とは、機械が明示的にプログラムされることなく、データから学習して課題を解決する能力と定義される）（著者翻訳）**出典：ENISA "Securing ML Algorithms** [3]**" ,2021 年**

Machine learning (ML) refers more specifically to the "field of study that gives computers the ability to learn without being explicitly programmed," or to computer programs that utilize data to learn and apply patterns or discern statistical relationships.
（機械学習（ML）とは、より具体的には、「コンピューターに対して、明示的にプログラムされることなく、学習する能力を与える研究分野」、あるいは、データを活用してパターンを学習・適用させ、統計的関係を識別するコンピュータプログラムである。）（著者翻訳）**出典：米国国立標準技術研究所 NIST Special Publication 1270 "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence** [11]**" ,2022 年**[7]

Machine learning systems are designed to learn patterns and associations from data. Typically, a machine learning method consists of a statistical model of the relationship between inputs and outputs, as well as a learning algorithm. The algorithm specifies how the model should change as it receives more information (in the form of data) about the input–output relationship it is meant to represent. This process of updating the model with more data is called "training."
（機械学習システムは、データからパターンや関連性を学習するために設計されている。一般に、機械学習の手法は、入力と出力の関係を表す統計モデルと、学習アルゴリズムで構成される。学習アルゴリズムは、入出力の関係についての情報（データ）を受け取ると、モデルをどのように変化させるかを指定する。より多くのデータによってモデルを更新するこのプロセスは、「学習」と呼ばれている。）（著者翻訳）**出典：Center for Security and Emerging Technology(CSET) "Key Concepts in AI Safety: Robustness and Adversarial Examples** [12]**" , 2022 年**

[7] *「コンピューターに対して、明示的にプログラムされることなく、学習する能力を与える研究分野」* の部分は、米国の計算機科学者アーサー・サミュエルによる言葉（1959 年）である。

上で見たとおり、機械学習の説明にはある程度の幅がある。本書においてはこれらの説明を参考とし、**「機械学習」**に対して「明示的なプログラムを与えられることなく、入力されたデータから、一定のアルゴリズム（計算方法）に基づいてコンピューターがパターンを発見し、新たに与えられたデータに対して発見したパターンを当てはめることで識別や予測等を可能とする手法」という説明を採用することとする。

なお、機械学習の一つの手法として、「ニューラルネットワーク」という手法がある。これは、脳内の神経細胞（ニューロン）に着想を得た数理モデルであり、入力に対して所定の変換をして出力する処理単位がネットワーク状に結合したものであり、図 2-1 のようなイメージで表される。

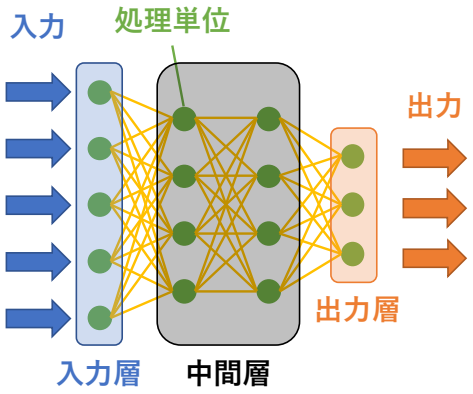

図 2-1 ニューラルネットワークのイメージ（黄線が結合を示す）

そして**「深層学習」**は「多数の層から成るニューラルネットワークを用いて行う機械学習」であるとされる［1］。すなわち、図 2-1 で中間層が多層となったニューラルネットワークが深層学習である。

深層学習は、「パターン/ルールを発見する上で何に着目するか（「特徴量」）を自ら抽出することが可能[8]［1］」であるのが大きな特徴である。

なお、AI、機械学習、深層学習の関係は図 2-2 のとおりである。

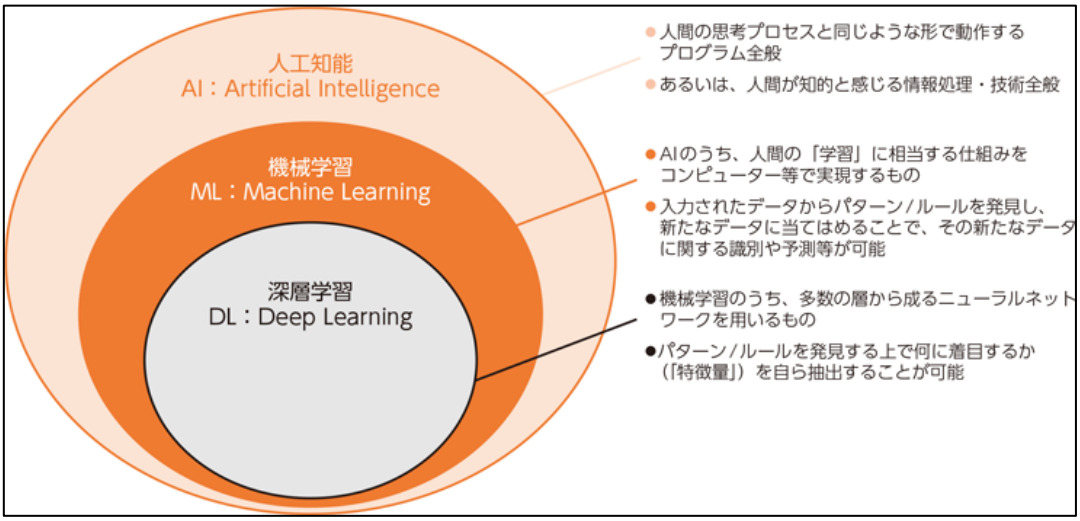

図 2-2 AI、機械学習、深層学習の関係（出典：総務省 「令和元年版 情報通信白書［1］」）

[8] ただし、実態は、膨大なデータから機械的・反復的な探索を行うことによって抽出している

2.3　学習と推論

2.2 節で述べた機械学習には大きく「**学習**」と「**推論**」の 2 つのプロセスがある［1］。

機械学習において「学習」とは、「明示的なプログラムを与えられることなく、入力されたデータから、一定のアルゴリズム（計算方法）に基づいてコンピューターがパターンを発見」するプロセスである。一定のアルゴリズム（計算方法）を「**機械学習アルゴリズム**」と呼び、入力するデータを「**学習用データ**」[9]と呼ぶ。

機械学習アルゴリズムが学習用データを学習することで、「**機械学習モデル**」が作られる。学習が行われたことを示すために「**学習済みモデル**」などとも呼ばれる。

一方「推論」とは、機械学習モデルが、「新たに与えられたデータに対して発見したパターンを当てはめることで識別や予測等を可能とする」プロセスである。識別や予測された結果を「**推論結果**[10]」と呼ぶ。学習されたパターンやルールが当てはめられる新たなデータを、本書では「**推論用データ**[11]」と呼ぶことにする。すなわち、機械学習においては、学習用データと推論用データという、使われ方の異なる二種類のデータがあることになる。以上をまとめると図 2-3 のようになる。

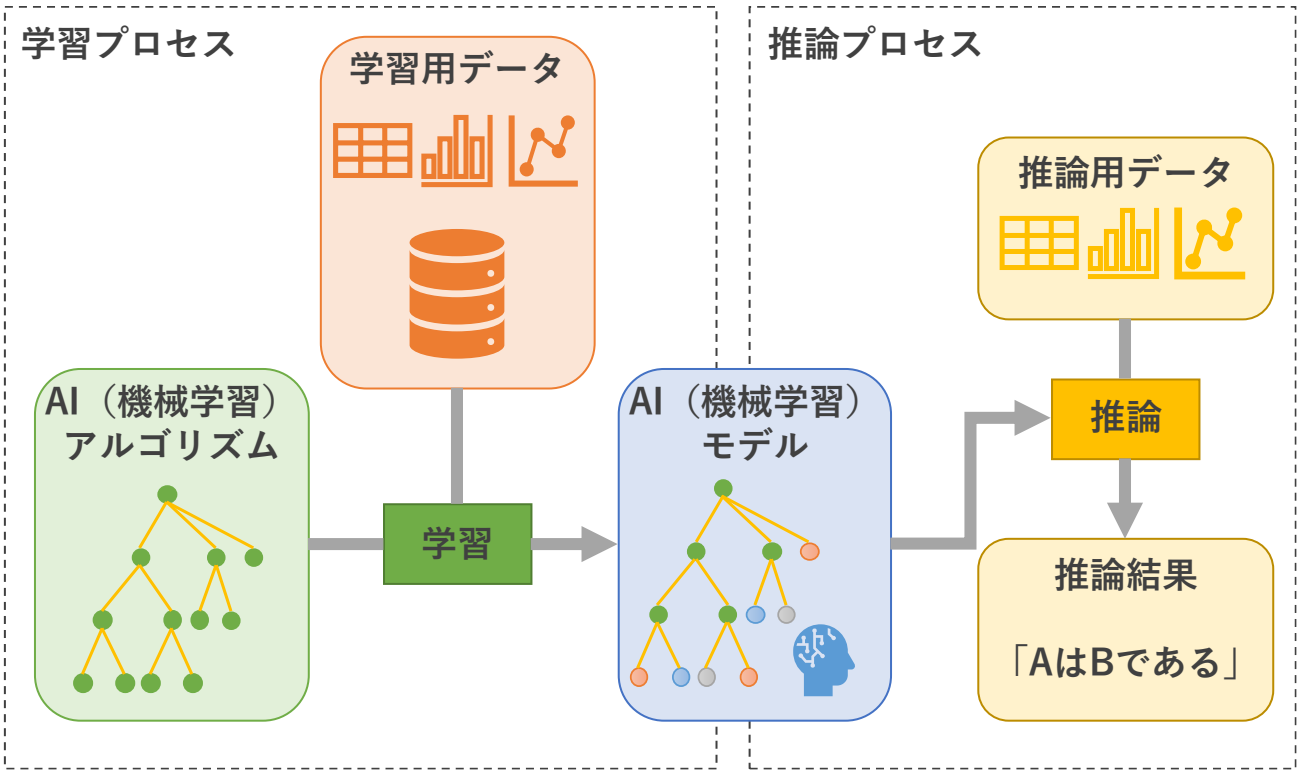

図 2-3 学習プロセスと推論プロセスのイメージ

[9] 学習用データは、訓練用データ、トレーニングデータなどとも呼ばれる

[10] 単に「出力」などと呼ばれる場合もあるが、推論プロセスでの出力であることを明確化するため「推論結果」を使う

[11] 単に「入力」「入力データ」などと呼ばれる場合もあるが、学習用データと区別するために「推論用データ」を使う

機械学習アルゴリズムは人間が作ったものであり、学習用データも人間が準備するものであり、学習は人間がコンピューターに実施させるものである。その意味では、これらは原理的には全て人間が管理・把握できるはずである。

しかし、学習の結果として作られた機械学習モデルは、学習・推論両面で処理が複雑で、学習・推論の挙動を人間が解釈できない場合がある。特に深層学習ではそれが顕著である。

それを人間から見ると、あたかも AI が「**ブラックボックス**」であるように見える。そのため「AI はブラックボックスである」と扱われていることもある［13]。

AI がブラックボックスであること自体が問題になる場合、ならない場合がそれぞれあり、状況に応じて適切に使うことが必要である［14]。また「説明可能 AI（5.4 節)」と呼ばれる、推論結果に対する判断根拠を説明できる AI の研究も進んでいる。

しかし、まだ説明可能 AI とは言えない AI の利用が多いため、**本書においては「AI はブラックボックスである」という立場で議論を進めていく。**

2.4　学習用データの質

機械学習では、学習用データを学習して、機械学習モデルが作られる。そのため、機械学習モデルの性能は学習用データの質に左右される。

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［15]」においては「データセットの被覆性」「データセットの均一性」「データの妥当性」が言及されている。

機械学習においては、解決したい課題に対して、例えば、

- ✓ 学習用データが対象とする範囲を十分にカバーしていない。例えば、動物を画像分類したい場合に、学習用データに哺乳類がほとんど含まれていない場合など。
- ✓ 学習用データに偏りがある。例えば、動物を画像分類したい場合に、学習用データの中で、猫の画像だけ、ほかの動物の画像に対して圧倒的に枚数が多い場合など。
- ✓ 正解ラベル（2.5.1 項）が誤っている。例えば、動物を画像分類したい場合に、学習用データの中の犬の画像に猫のラベルがつけられている場合など。

などが学習用データの良くない例として考えられる。学習用データがこうしたデータとならないように注意する必要がある。

2.5　機械学習の分類

機械学習には、大きく「**教師あり学習**」「**教師なし学習**」「**強化学習**」という 3 つの手法がある［1］[12]。これらの手法はどれかが一方的に優れているというものではなく、特性に応じて使い分けられている。

2.5.1　教師あり学習

教師あり学習とは、「**正解ラベル**」をつけた「**正解データ**[13]」を含むデータで学習をさせる方法である。

教師あり学習には、例えば、画像の「分類」などがある。

例として、「これはリンゴの画像である」ことを示すラベル（正解ラベル）をつけたリンゴの画像を「正解データ」とし、これを含む学習用データで学習を行うことで、ある画像がリンゴであるかどうかを推論する AI ができる。

「分類」のほか、教師あり学習は学習用データの傾向をもとに予測を行う「回帰」にも用いられる。

教師あり学習には教師なし学習に比べ精度の高い推論が可能であるというメリットがある。一方で、大量のデータに正解ラベルをつける手間がかかる、などのデメリットがある。

2.5.2　教師なし学習

教師なし学習とは、正解データを含まないデータで学習をさせる方法である。

教師なし学習は、例えば「クラスタリング」などがある。クラスタリングとは、データを似た性質を持つグループに分けることである。

例として、購買状況などのデータをもとにした顧客のグループ化が考えられる。「この顧客は○◯グループである」といった正解を与えているわけではないので、推論の結果得られた各グループの意味を AI 自体が明らかにすることはできないが、人間が適宜解釈を行えば意味のあるグループ分けができる。

教師なし学習には、正解データにラベルをつける手間が不要である、正解のない分野でも使うことができるというメリットがある。一方で、教師あり学習に比べ推論の精度で見劣りする、分類した各グループの意味を人間が解釈する必要があるなどのデメリットもある。

2.5.3　強化学習

強化学習とは、ある環境の中で、AI 自身が、求める成果[14]を最大化するように学習の試行錯誤を繰り返す方法である。試行錯誤の過程で得られるデータなどが学習用データとなる。

将棋や囲碁の最適手を提案する AI として使われ、成果を挙げている［1］。

---

[12] ただし、「半教師あり学習」などもあり、3 つに限定されるものではない

[13] 正解データは教師データとも呼ばれる

[14] 「報酬」と呼ばれる

2.6　AI システムの企画・開発・運用プロセス

AI システムの企画・開発・運用には様々なプロセスが存在する。それらのプロセスは目的や組織により様々に整理されている。

例えば、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［16］」では次のように整理している。

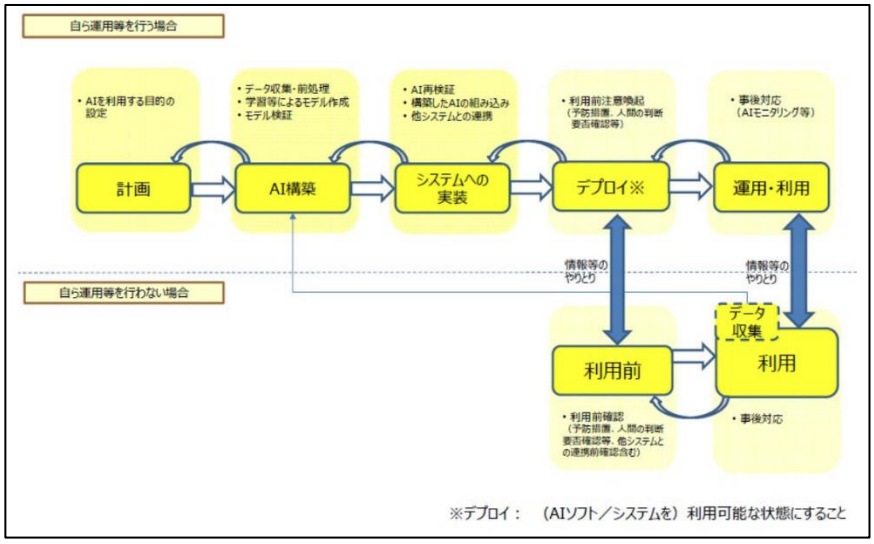

図 2-4「AI 利活用ガイドライン」におけるプロセス

また、PwC グループでは次のように整理している[15]。

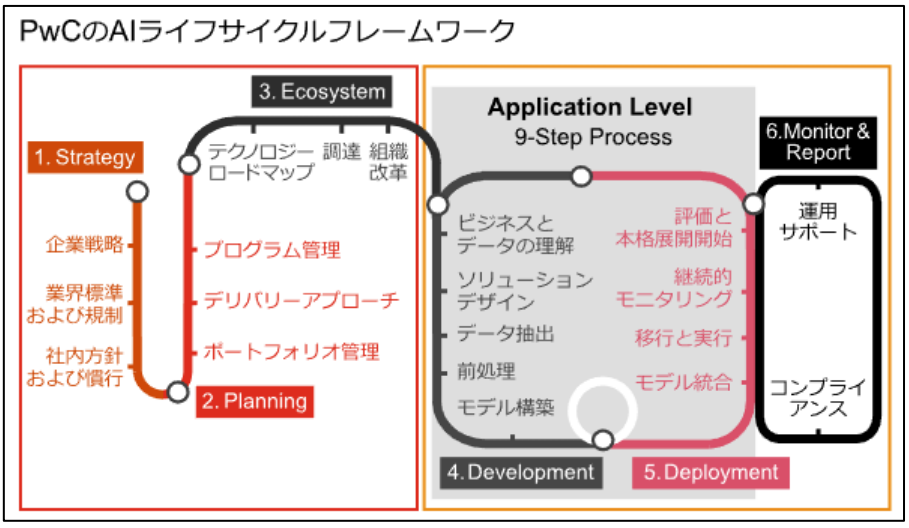

図 2-5 PwC グループにおける AI ライフサイクルフレームワーク

その他の組織においても、自組織の活動に紐づけて AI システムの企画・開発・運用プロセスを策定、公開している[16]。それらの多くに共通しているのは次の点である。

---

[15] PwC Japan<https://www.pwc.com/jp/ja/services/consulting/analytics/responsible-ai.html>

[16] 例えば、伊藤忠テクノソリューションズ社
<https://www.ctc-g.co.jp/bestengine/article/2017/1120a.html>

✓ ビジネス目標や解決したい課題、AI 利用の目的を明確化するプロセスが存在する
✓ データを収集し、AI の学習用に前処理するプロセスが存在する
✓ 開発した AI を利用して終わりではなく、AI の精度などをモニタリングし、必要に応じて**再学習**を行うなど、プロセスのループが存在する

本書では、上の内容と第 4 章の内容を考慮し、AI システムの企画・開発・運用プロセスを次のように整理して扱う。プロセスのループを含めていることから「**AI ライフサイクル**」と呼ぶこととする（表 2-1、図 2-6）。

表 2-1 本書における AI システムの企画・開発・運用プロセス（AI ライフサイクル）

プロセス	区分	概要
課題設定	企画	AI で解決したい課題を明確にする。課題を解決するための AI の種類を設定したり、求められる AI の精度などを設定したりする。
データ収集	開発	AI 構築のために必要なデータを収集する。収集するデータは、組織内データ、公開データ、他組織が提供するデータ、などがある。
データ前処理	開発	収集したデータを AI 構築に適した形に変換する。ファイルフォーマットの変換、異常値の除去、ノイズの除去、データの匿名化・仮名化、学習用データの補強（データ拡張）などが含まれる。
AI 構築	開発	前処理されたデータをもとに AI を構築する。AI（機械学習）アルゴリズムの選択を行い、学習用データを学習し、AI（機械学習）モデルを構築する。
AI 運用	運用	構築された AI を利用する。AI を利用したい環境にデプロイし、AI に推論させたいデータ（推論用データ）を入力して推論させて結果を得る。推論の結果をモニタリングし、時間経過による精度低下[17]などを検知したら、必要に応じて AI のメンテナンスを行う。

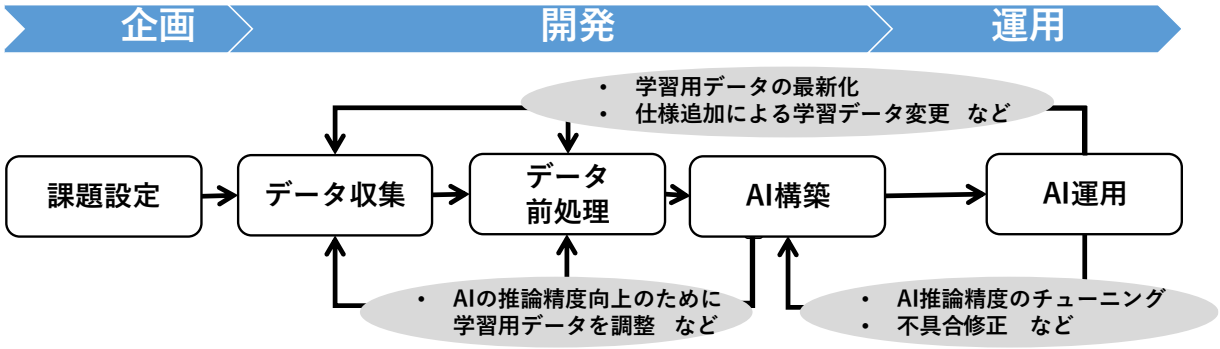

図 2-6 本書における AI システムの企画・開発・運用プロセス（AI ライフサイクル）

[17] 時間経過による精度低下を「ドリフト」という。学習用データと推論用データが異なる分布となる「データドリフト」や、正解ラベルの解釈が学習時と変化する「概念ドリフト」などがある。
例えば、次の解説が参考になる。
<https://atmarkit.itmedia.co.jp/ait/articles/2202/21/news033.html>

2.7　従来のシステム開発と AI システム開発の違い

AI が組み込まれたシステムの開発においては、AI が組み込まれていない従来のシステム開発に比べて、例えば次のような差異があると考えられる [17]。

- ✓ AI 開発には従来のシステム開発とは異なる専門的な知識を必要とする。システムの開発と AI 部分の開発でベンダーが異なる場合もある（図 2-7 上）。
- ✓ AI の開発には大量の学習用データを必要とする。学習用データは、自組織で準備する場合のほかに、外部から取得する場合、外部に作成を委託する場合などがある（図 2-7 下）。
- ✓ 教師あり学習の場合、学習用データに正解ラベルを付ける作業が発生する。自組織だけで対応できず、外部に作業を委託する場合がある（図 2-7 下）。

このため、AI の開発では関係者が多数となる場合があり、権利関係が複雑になる、セキュリティ水準にばらつきがある、悪意を持った関係者が紛れ込む、などの点に注意が必要となる。

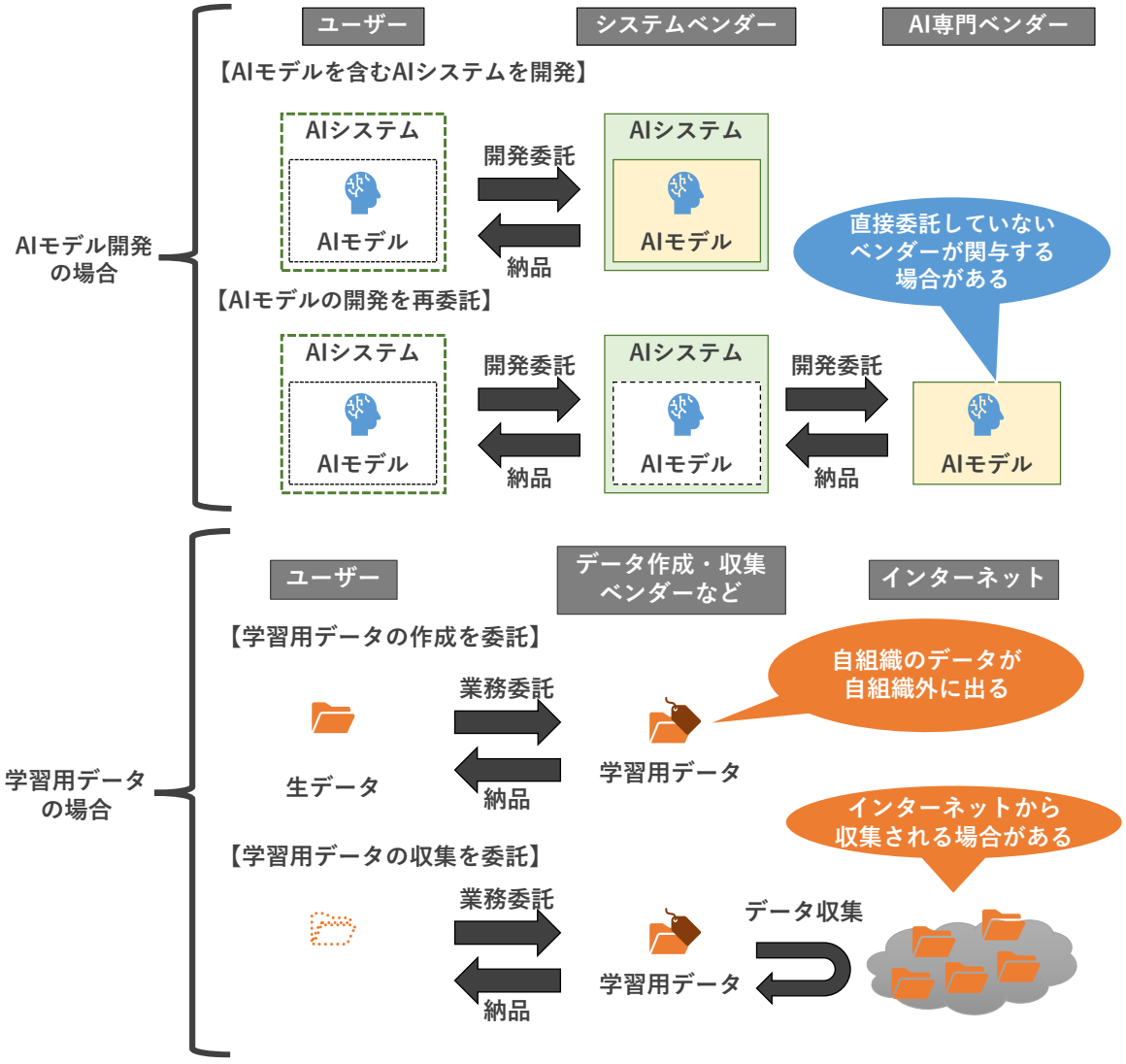

図 2-7AI モデル開発、学習用データ準備の類型と注意点（吹き出し部分）

2.8　公開されているライブラリ、モデル、データなどの利用

インターネット上には、次に示すような各種のライブラリ、モデル、データが公開されている。

AI 開発においては、開発効率を高めるために、これらを利用もしくは再利用する場合も多い。

ただし、これらの公開されているライブラリ、モデル、データの信頼性は公開元による。すなわち、全てが信頼できるわけではない点に注意する必要がある。例えば、マルウェアの感染リスクやバックドアなどの悪意あるプログラムが仕込まれている（4.7.1 項）などの可能性が否定できない。また、公開されているものであっても、ライセンスが設定されている場合などもあるため、再利用にあたっては注意が必要である。

2.8.1　機械学習ライブラリ

AI の開発においては、機械学習モデルを構築するためのプログラムをまとめた「**機械学習ライブラリ**[18]」というものがよく利用される。有名なものとしては、オープンソースの「scikit-learn[19]」、深層学習用に特化した Google 社の「TensorFlow[20]」、オープンソースの「PyTorch[21]」などがある。

他にも、個人や特定の組織が作成した機械学習ライブラリが多数あり、「GitHub[22]」などで公開されている。

2.8.2　学習済みモデル

学習済みモデルが格納されたファイルなどがインターネット上に公開[23]されており、それを他者が取得して再利用することがある。

2.8.3　公開データ

インターネット上では、AI で利用可能なデータが公開[24]されており、学習用データや推論用データなどに使うことができる。官公庁、大学、企業や個人が公開している場合もある。

2.8.4　データ収集ツール

これらのデータをインターネット上から取得する際に、クローリングやスクレイピングのツールやライブラリ[25]を使う場合がある。特定の Web サイトなどに特化したツールやライブラリも存在する。

---

18 「ライブラリ」と「フレームワーク」は厳密には異なる概念であるが、本書では「ライブラリ」に「フレームワーク」を含むものとする

19 <https://scikit-learn.org/>

20 <https://www.tensorflow.org/>

21 <https://pytorch.org/>

22 <https://github.co.jp/>

23 例えば、「Hugging Face Hub」<https://huggingface.co/models> など

24 例えば、リクルート社 <https://www.megagon.ai/jp/projects/datasets/> など

25 例えば、「Scrapy」<https://scrapy.org/> など

2.9 AI 活用事例

本項では、AI の活用事例を紹介する。なお、これらの事例は第 4 章においても事例として触れる。

2.9.1 スマートスピーカー[26]

スマートスピーカーは、話しかけることで、録音、音楽再生、照明や家電のスイッチの切り替え、インターネットを経由した情報の取得および音声による通知を行う商品・サービスの総称である。話しかけた声の認識とテキスト化（音声認識）、テキスト内容の処理（自然言語処理）、処理すべき内容の判断などに AI が活用されている。

スマートスピーカーは今後も、様々なサービスと連携することで、例えば物品の購入やホテル・タクシーの予約など、より活用範囲が広がることが想定される。

2.9.2 与信管理[27]

金融業では、リスク管理のために個人・企業の与信管理を行っている。従来は、取引状況や決算状況などのデータからルールベースの処理によって変化を予測し、問題の兆候があれば対応する、という対策をとっていたが精度向上に課題があった。変化の予測に AI を活用することで、膨大なデータを処理し、将来の状況や業績を精度良く予測できるようになり、より正確なリスク管理が可能となる。

2.9.3 品質管理[28]

AI の典型的な活用例として画像分類がある。これは、大量の画像データ（学習用データ）から AI が画像の特徴を自動で学習し、新たな画像を特徴に応じて分類することを可能にするものである。画像分類の製造業における応用として、品質管理がある。これは、品質検査工程で画像分類の技術を用いて、加工した製品が良品か不良品かを判定する、というものである。高精度な AI を用いると、人間が目視で行う精度を超え、業務を代替できる場合がある。

---

[26] 例えば、消費者庁「AI 利活用ハンドブック」P11 など
<https://www.caa.go.jp/policies/policy/consumer_policy/meeting_materials/review_meeting_004/assets/ai_handbook_200804_0002.pdf>

[27] 例えば、日本銀行 金融高度化センターワークショップ資料 など
<https://www.boj.or.jp/announcements/release_2019/data/rel190215d2.pdf>

[28] 例えば、経済産業省「AI 導入ガイドブック 外観検査（部品、不良品あり）」など
<https://www.meti.go.jp/policy/it_policy/jinzai/AIguidebook_gaikan_furyo_FIX.pdf>

2.9.4　製造プロセスの置き換え[29]

工作機械は、加工プロセスのために高精度な制御や調整を必要とする。温度や湿度、機械の振動などの環境や、工作対象の材質、表面の状況など様々な要因を考慮して制御を行う必要があり、ルールベースの処理では対応が難しい。そのため、作業員が工作機械を手動で制御、調整している。

工作機械などに IoT センサーを取り付け、大量のデータを取得し、AI で処理させることで、高精度な制御や調整が可能となる。これにより、熟練工の退職による技術継承の断絶などの課題に対応できるようになる。

2.9.5　機器の障害・寿命予測[30]

鉄道、電気、ガス、水道などの社会インフラを支える機器類は、使用される期間が長くなればなるほど劣化が進み、破損・故障のリスクにさらされることになる。この予期せぬ破損や故障は社会生活に必要不可欠なサービスの供給停止のみならず、時にサービス利用者の死傷事故につながるおそれもあることから、法令で定められた定期的な点検の結果や過去の実績から算出された耐用年数での予防保全的な交換が行われている。しかしながら、点検周期の合間での故障や、コンディションの比較的良い機器も一律に予防交換しなければならないなど、必ずしも最適とは言えない面もあった。

近年、こうした機器に振動や音響などのセンサーを取り付け、故障に至るまでの詳細なデータを収集し、AI に処理させることで、故障の予兆をより正確かつリアルタイムに予測する試みが行われている。個々の機器の実際の状態に即した交換の実施など、より合理的な運用につながることが期待されている。

2.9.6　AI を活用した Web アプリケーション診断[31]

Web アプリケーション診断を行う際は、診断対象の Web アプリケーションを隅々までクローリングするなどの作業が必要である。従来はそれらの作業を手作業で行っていたが、これは Web アプリケーションの規模によっては非常に時間のかかる作業である。AI を活用することで、ページ種別を判断し、Web アプリケーションに適切なパラメータを入力し、自動的にクローリングを行えるようになるため、アプリケーション診断の効率化が期待される。

ページ種別の認識や、ページ遷移の成否にはテキスト分類に用いられる AI を用いて実装し、入力フォームに最適なパラメータ値を入力する際には、画像認識などで使用されている教師あり学習や強化学習が用いている例がある。

---

[29] 例えば、経済産業省 中部経済産業局 電力・ガス事業北陸支局「課題解決のための IoT・AI 活用ガイド」P20 など <https://www.chubu.meti.go.jp/e21shinsangyo/190403/guide.pdf>

[30] 例えば、経済産業省 産業保安グループ「スマート保安の促進～産業保安分野におけるテクノロジー化の推進～」P2 など
<https://www.meti.go.jp/shingikai/sankoshin/hoan_shohi/sangyo_hoan_kihon/pdf/002_01_00.pdf>

[31] 例えば、三井物産セキュアディレクション「機械学習を用いた診断 AI の概要」など
<https://www.mbsd.jp/blog/20160113.html>

第3章　　AI を活用したサイバーセキュリティ対策

近年、サイバー攻撃はますます高度化してきており、従来の「ルールベース」「シグネチャベース」のセキュリティ対策による対処が難しくなってきている。

サイバーセキュリティ対策に AI を活用することによって、膨大なデータを処理可能となり、また、ルールベース・シグネチャベースで発見できなかった脅威（ゼロデイ攻撃など）を防御できる可能性がある［10]。

本章では、AI を活用したセキュリティ製品やサービス（以降、「AI 活用セキュリティ製品」）として著者が Web 調査したものを示し、それらを企画・導入・運用する際の留意点について述べる。

AI 活用セキュリティ製品には大きなメリットがある一方、企画・導入・運用にあたっては AI の特徴に基づいて留意すべき点がある。3.2 節では、その留意点を説明する。

なお、ユーザー企業において、AI 活用セキュリティ製品を利用する場合、ユーザー企業が AI を開発する場合は少なく、セキュリティベンダーが開発した AI を利用する場合が多いと考えられる。本章ではその状況を前提に留意点を検討している。

ただし、そうした場合であっても、ユーザー企業側のログデータなどが学習用データ、推論用データとして使用されている。すなわち AI 活用セキュリティ製品に学習用データや推論用データとして入力される、もしくはセキュリティベンダーに提供されている。

そのため、留意点をお読みいただく上で、第 2 章で述べた AI 知識を理解していることが望ましい。

留意点の検討にあたっては、「AI 活用セキュリティ製品は AI のセキュリティ分野への応用である」という観点も取り入れ、「G20 AI 原則［4]」、「人間中心の AI 社会原則［5]」や「AI 原則実践のためのガバナンス・ガイドライン ver. 1.1 [6]」も参考にしている。

また、本章と同趣旨の検討を行っている、英国 National Cyber Security Centre（NCSC）の Web ガイダンス”Intelligent security tools [7]”も参考にしている。

ただし、本書はこれらの資料の内容の網羅や準拠を意図しているものではない。

3.1 AI を活用したセキュリティ製品・サービスの種類

AI 活用セキュリティ製品の種類には、例えば、次のようなものがある[32]。

全ての製品を調査しきれているわけではないため、ここに示したものはあくまでも一例である。技術の進歩とともに AI を活用したセキュリティ製品の種類も増えてくるものと思われる。

なお、例えば、AI（機械学習）を使っていない IDS・IPS や、機械学習ではない AI を使っているソースコード診断製品などもある。表 3-1 に掲載されている製品・サービスが必ずしも機械学習の仕組みを搭載しているわけではない。

表 3-1 AI 活用セキュリティ製品の種類の例

製品・サービス[33]	主な適用対象	防御・検知できる対象
NGAV	PC、サーバ	マルウェアの振る舞いなど
EDR	PC、サーバ	マルウェアの振る舞いなど
WAF	公開サーバ	不正通信、DDoS 攻撃など
IDS・IPS	ネットワーク、サーバ	不正通信など
NGFW	ネットワーク、サーバ	不正通信など
SIEM・UEBA	各種ログ	被害の横展開、内部不正など
ブラウザ Web フィルター	PC	フィッシングサイトなど
メールフィルタ	メール	フィッシングメールなど
CAPTCHA	公開サーバ	Bot による不正ログインなど
AI 活用ペネトレーションテスト	システム	システムの脆弱性など
ソースコード診断	ソースコード	脆弱性を生むソースコードなど

これらの製品は、何らかのファイルやログを学習用データ、推論用データとして用いるものが多い。

例えば、NGAV はファイルを推論用データとして使用し、マルウェアなどを検知する。WAF や IDS・IPS などでは正常通信のログを学習用データとして使用し、それ以降の運用時の通信ログを推論用データとして使用し、不正な通信を検知する。

SIEM や UEBA でも、正常時のログを学習用データとして使用し、それ以降の運用時の通信ログを推論用データとして使用し、不審な振る舞いを検知する。

AI 活用ペネトレーションテストでは、Web のソースコードやリクエストに対するレスポンスを推論用データとして使用する。ソースコード診断ではソースコード自体が推論用データとなる。

AI 活用セキュリティ製品を活用するためには、学習用データ、推論用データを AI 活用セキュリティ製品に入力し、AI を自組織に適合させるための事前準備や運用が重要となる。以降の留意点でそれらを説明する。

---

[32] 本書著者による調査。

[33] 各ソリューションの略称については、用語集参照。

3.2　AI 活用セキュリティ製品の企画・導入・運用上の留意点

本節では AI 活用セキュリティ製品の企画・導入・運用上の留意点について、図 3-1 のように「企画」「導入」「運用」「共通」のプロセスに区分した上で説明する。

なお、理解の助けとなるように説明中に簡潔な例を挙げている。さらに、他の例を【〇〇】の形で強調しページの下部に掲載した。ただし、それに限定されるものではなく、各組織の状況に応じて検討を行っていただきたい。

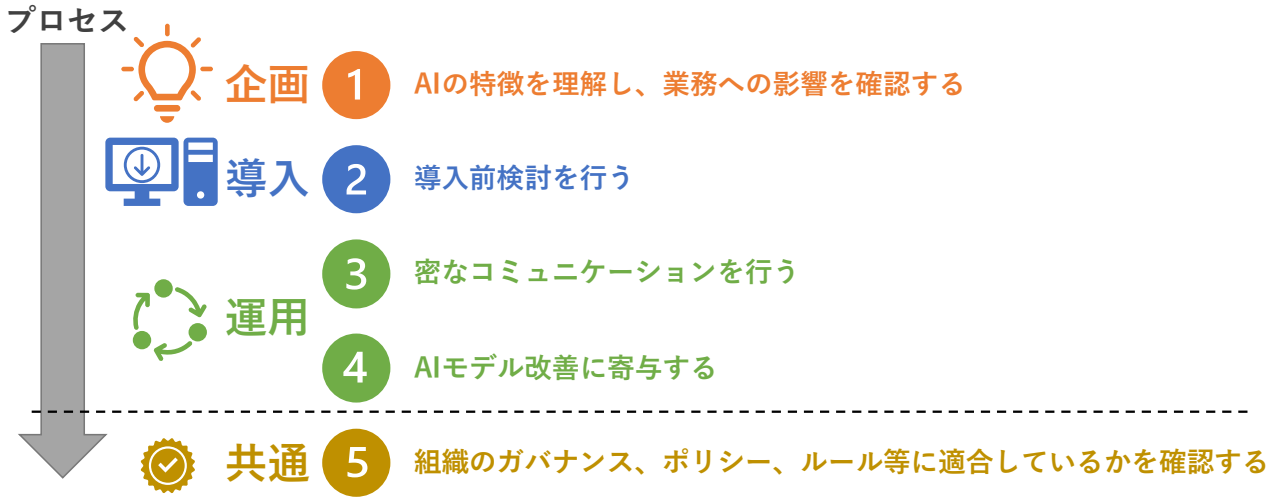

図 3-1 AI 活用セキュリティ製品の企画・導入・運用プロセス

また、留意点は「AI 活用セキュリティ製品特有の留意点」を中心に述べているが、図 3-2 にある一般的な留意点をも一部含んでいる。図 3-2 の赤い部分全体が本章のスコープとなる。

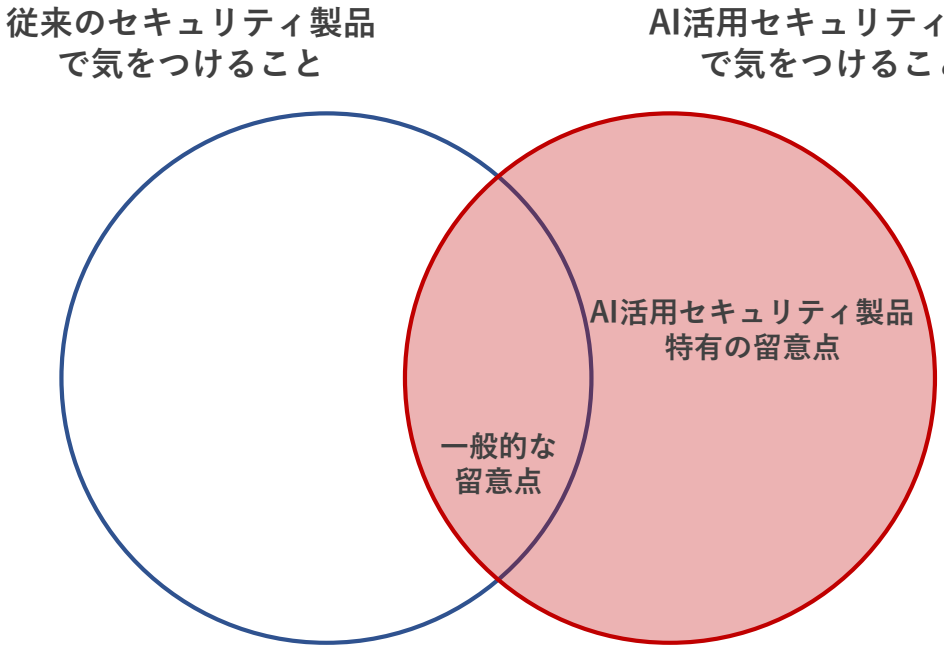

図 3-2 本章のスコープ

➢ **留意点１：AI の特徴を理解し、業務への影響を確認する**

第 2 章で述べたような AI の性質が、AI 活用セキュリティ製品においては次のようなメリット、デメリットとして反映される。

＜メリット＞

・人間が個別のルールを詳細に検討、実装することなく運用が可能である
・ゼロデイ攻撃などの未知の攻撃にも対応できる可能性がある
・大量かつ複雑な、各種セキュリティ製品のログを相関分析できる
・データをもとに AI モデルが再学習（更新）され、多くの場合、性能が向上する

＜デメリット＞

・「過検知」（例：正常な対象が誤って不正な対象として検知される）が発生する
・「すり抜け」（例：不正な対象として検知すべきものを見逃す）が発生する
・検知や動作の理由を説明できない場合がある
・データをもとに AI モデルが再学習（更新）され、事象の再現性がなくなる

また、AI 活用セキュリティ製品を適切かつ有効に活用するために、セキュリティベンダーなどから製品に関する情報を提供してもらうことが望ましい。

例えば、**AI モデルのアップデートの連絡有無、頻度、適用方法（自動・手動）**などを確認する。留意点 3 で述べるコミュニケーションにとって重要な情報である。

他にも、AI セキュリティ製品の**過検知に対する代替策**として、許可リスト機能が搭載されているかを確認する。搭載されていない場合、過検知に伴う正常ブロックを解消できないリスクが考えられる。許可リスト機能が搭載されている場合でも、例えば IP アドレスしか許可リスト登録できず、特定の文字列などが登録できない場合もある。自分たちの要件に応じた機能となっているかを確認する。

一部の産業用システムなど、通信要件が厳格に定められており、かつ過検知が許容できないシステムの場合には、ルールベースのセキュリティ製品による防御の方が有効である場合もある。AI 活用セキュリティ製品の適用が常に適切であるとは限らない。

業務や事業に与える影響をよく確認し、適切なセキュリティ製品を選択してほしい。

➢ **留意点２：導入前検討を行う**

AI 活用セキュリティ製品には、自組織のデータを学習用データとして学習し、学習した AI モデルで推論を行う製品がある。例えば、SIEM や UEBA[34]では組織の多数のログを連携させ、普段の通信などを学習する。そして、その学習用データから AI モデルを構築し、検知を行う。

そうした製品の場合、仕様による確認だけではなく、採用しようとしている AI 活用セキュリティ製品が、自組織のデータにおいても目的に応じた性能を出すかを、試行ライセンスで実際の製品を使って確認することがより望ましい。

EDR や WAF などの製品においては、「検知モード」「ブロックモード」が設定できるものがある。「検知モード」で普段の振る舞いや通信などを学習し、そこで作られた AI モデルをもとに「ブロックモード」で不正な振る舞いや通信を実際にブロックするというものである。

AI が高い性能を出すためには、ある程度のデータ量、もしくは学習期間が必要である。性能を適切に測るためには、データの準備、データ量や学習期間の観点を十分に意識し、導入前検討に臨むことが必要である。

AI 活用セキュリティ製品の試行ライセンスの期間は限られているため、**【製品の検証】を行う前に、**導入前検討に必要な事項を**【事前確認】**して準備しておくことが望ましい。

【事前確認】の例

✓ 必要なデータの準備が揃っているか？
✓ データ量や学習期間は十分か？
✓ 導入前検討に必要な自組織のデータが準備できているか？
✓ データ連携を行う場合、その仕組みは準備できているか？
✓ データ連携を行う際の、ネットワーク負荷は大丈夫か？

【製品の検証】の例

✓ AI 活用セキュリティ製品の性能を出すための自組織のデータは質、量ともに十分か？
✓ データや製品間の連携が実際にできるか？
✓ 自組織のデータにおける「過検知」や「すり抜け」の頻度は許容範囲内か？
✓ アラートやサービス画面から出力される情報は十分か？

検討の過程は、**事象の再現性がない場合に備え**て、適宜メモやスクリーンキャプチャ、などでもよいので、何らかの形で記録を取っておくことが望ましい。なお、これらの記録は試行ライセンスの期間満了後の資料作成などにも利用できる。

[34] UEBA: User and Entity Behavior Analysis の略称。ユーザーや機器の振る舞いを分析するソリューション。

➢ **留意点３：密なコミュニケーションを行う**

AI 活用セキュリティ製品には AI モデルの更新を頻繁に行う、もしくは随時更新を行う[35]場合があり、業務に影響を与える可能性がある。

従来のセキュリティ製品のバージョンリプレース時には、ベンダーと業務影響を相談の上で慎重に進めるが、AI セキュリティ製品では頻繁に AI モデル更新が行われることがあり、**意図しない影響が発生する可能性が従来製品より高い**。そのため、日頃からセキュリティベンダー、システムベンダーなどと密なコミュニケーション体制を構築していることが望ましい。

また、AI モデルが更新される場合、これまで検知・ブロック対象外であったプログラムが、突然検知・ブロックの対象となる場合がある。AI モデルを自分たちで更新する場合、セキュリティベンダーが更新する場合、いずれもスケジュールや状況を把握してシステム担当者などにも共有することで、問題が起きた際の早期対応に役立てることができる。

このように、AI 活用セキュリティ製品をうまく運用していく上で、組織内やセキュリティベンダー、システムベンダーなどの関係者とのコミュニケーションは重要である。

[35] 例えば、オンライン学習。人工知能学会 <https://www.ai-gakkai.or.jp/resource/my-bookmark/my-bookmark_vol30-no5/> による解説が参考になる

➢ 留意点４：AI モデル改善に寄与する

留意点１で、AI 活用セキュリティ製品は、「人間が個別のルールを詳細に検討、実装することなく運用が可能である」「データをもとに AI モデルが再学習（更新）され、基本的には、性能が向上する」ことがメリットであると述べた。一方、**ユーザー側で AI モデルがチューニング可能**な場合[36]もある。その場合、自組織の状況に応じて**チューニングを行うことで AI モデルの精度を向上できる可能性**がある。

また、AI 活用セキュリティ製品は、製品ごとに独自の AI モデルを用いてサイバー攻撃などを検知しているが、どうしても「過検知」や「すり抜け」が発生する。

別の手段、例えば多層防御の後段のセキュリティ製品などで、「**過検知」や「すり抜け」を発見した場合、その情報や検体を AI モデル運営者に提供することで AI モデルの改善に寄与**してほしい。

例えば、AI メールフィルタをすり抜けてきたスパムメールなどは、開封せずに AI メールフィルタ運営ベンダーに報告する。AI メールフィルタ運営ベンダーはその情報をもとに AI モデルを改善させることができる。

他にも、AI 活用セキュリティ製品からアラートが発報された場合に、その**アラートに対して、正解・不正解をフィードバックすることで、AI モデルが再学習され、AI モデルの精度が向上する**場合[37]がある。

第 2 章で述べたとおり、AI モデルは学習用データからの学習によって構築される。しかし、セキュリティの分野において、「過検知」や「すり抜け」が起きる推論用データは少ない。

なぜなら、例えば、組織のネットワークで流れるトラフィックのほとんどは正常な通信であり、セキュリティ製品はその中のごくわずかな不正通信を検知するためである。つまり、不正の正解ラベルがついたデータは非常に少なく、学習用データが正常なデータに偏った状態である場合が多い。

「過検知」や「すり抜け」が起きた推論用データを AI モデル運営者に提供することは、学習用データの偏りの減少に貢献し、AI モデルの精度向上に寄与する。

---

[36] 例えば、キャノンマーケティングジャパン「ソリューション理解から製品比較のポイントまで」<https://eset-info.canon-its.jp/files/user/pdf/business/threat-solution/paper/wp202104_MEEI2104.pdf>

[37] 例えば、OKI テクニカルレビュー 2018 年 12 月/ 第 232 号 Vol 85 No.2 <https://www.oki.com/jp/otr/2018/n232/pdf/otr232_r13.pdf>

➢ **留意点５：組織のガバナンス、ポリシー、ルール等に適合しているかを確認する**

AI 活用セキュリティ製品の企画・導入・運用が、組織が定めているガバナンス、ポリシー、ルールに適合しているかを確認することが望ましい。

例えば、従来のセキュリティ製品においても「管理者や特権ユーザーの監査」を行っている組織は多い。セキュリティガバナンスの観点から、それは適切である。

ところで、AI 活用セキュリティ製品においては、**特に大量かつ複数種類のデータを処理する**場合が多い。例えば、Web アクセス履歴、ファイルアクセス履歴、メール送受信などの**データを組み合わせて分析することで高い検知率を実現**する製品も存在する。ただし、それらの**データを組み合わせることが、個人の行動特定や、機密情報の特定につながる**こともある。

そのため、AI 活用セキュリティ製品における「管理者や特権ユーザーの監査」は、従来のセキュリティ製品以上に重要となる。また、単純な AI 活用セキュリティ製品の管理画面へのログイン履歴だけでなく、どのようなデータにアクセスしたか、操作をしたか、などの確認も重要となる。

また、自組織が AI ガバナンスに関するポリシーやルールを定めている場合には、AI 活用セキュリティ製品とその運用がそれらに適合していることを求められる場合がある。先程の AI 活用セキュリティ製品の「管理者や特権ユーザーの監査」は **AI ガバナンス（5.2 節）の観点でも重要**である。

このほか、AI ガバナンスや AI 倫理（5.3 節）の観点から、AI 活用セキュリティ製品を適切に扱うことも求められる。例えば、**性別や健康状態などの情報は学習用データから除外する**ことが挙げられる。そうした情報を学習用データとして除外しない場合、AI 活用セキュリティ製品が差別や偏見につながる推論（出力）を行う場合がある。

3.3　AI 活用セキュリティ製品の企画・導入・運用上の留意点の想定事例

本節では、AI 活用セキュリティ製品の企画・導入・運用の想定事例をもとに、留意点がどのように反映されているかを述べる。**この想定事例は架空企業におけるフィクションである。**

【想定事例：WAF の導入】

ある企業（A 社）では、新規事業部門が、新規事業の展開のため、社外に各種の顧客向けシステム（以降、それらをまとめて、新規事業システム）を公開しており、ルールベースの WAF を導入している。新規事業の拡大とともに新規事業システムの数も増え、セキュリティ部門の WAF 運用の負荷が上がってきている。また、サイバー攻撃のリスクも懸念している。

あるとき、セキュリティベンダーから AI を活用した WAF（以降、AI-WAF）を紹介された。AI-WAF は AI を活用しており、過検知のデメリットは排除できないものの、「自動でルール設定され、ルール設定の負荷が少ない」「ゼロデイ攻撃のリスク低減」など、メリットがあることがわかった。また、許可リスト機能が搭載されていることもわかった。A 社は、新規事業のサービスレベルなども検討しながら、AI-WAF の導入を検討することにした（**留意点 1**）。

まず、新規事業部門に対して AI-WAF 導入の打診を行った。新規事業システムのベンダーが「システム仕様上のある特定通信の中に不正通信に似た文字列が含まれている。他社で WAF を導入した際に、その通信が過検知されてブロックされ、業務が止まった」点を懸念していることを把握した（**留意点 1**）。

新規事業システムに問題が発生すると新規事業の信用低下を招く。スケジュールやデータ量を確認の上、実際の通信やデータを対象として導入前検討（概念検証；PoC）を行うこととした。

PoC の結果、運用負荷が低減すること、許可リスト機能により特定通信の過検知を回避できること、などが確認できた（**留意点 2**）。この検証結果をもとに AI-WAF の正式導入を決定した。

AI-WAF の AI モデルは頻繁に更新されるが、セキュリティベンダー、A 社セキュリティ部門、A 社新規事業部門、新規事業システムベンダーの 4 者で AI モデルの更新タイミングや許可リスト登録の運用方法などを共有することに合意した（**留意点 3**）。なお、定期的に監査ログを新規事業部門に提供するなど、新規事業部門が定めるビジネスルールにも適合することができている（**留意点 5**）。

運用初期に、（特定通信ではない）正常通信でごくわずかに過検知によるブロックが発生したが、過検知したトラフィックを正常通信として登録し、AI に学習させることで、過検知はほとんど起きなくなった（**留意点 4**）。

ある日、ゼロデイ攻撃が発生し、同業他社が被害にあった。報道の直後、新規事業を展開している部門から被害に関する問い合わせがあった。AI-WAF のログを確認すると、ゼロデイ攻撃を防げていたことが分かった。AI-WAF は新規事業システムを適切に保護することで新規事業の拡大に貢献した。

第4章　　AI のセキュリティ

4.1　AI への攻撃に関する動向、事例

AI に対するサイバー攻撃は、活発に研究されている[38]。画像に特定のノイズを乗せることにより画像を誤認識させる（4.7.2 項）、チャットボットに差別的な言葉を学習させて差別的な発言をさせる[39]、攻撃者のコードにより AI 活用セキュリティ製品による検知の回避を許してしまうなどの事例（コラム参照）が報告されている。

AI に対するサイバー攻撃はすでに観測されており、今後さらに活発になると考えられる[40]。社会における AI のさらなる活用が進むことで、攻撃対象となる AI の増加や攻撃ツールの普及も進み、AI アルゴリズム・AI モデルに対する攻撃が現実的なものとなってくる。

AI のセキュリティを考えるとき、第 2 章で述べた「データを学習する」という AI の性質や、AI アルゴリズムと AI システムの開発・提供ベンダーが異なる（2.7 節）、公開データや他社提供データを使う（2.8 節）、オープンソースの機械学習ライブラリを活用する（2.8 節）、などの状況から、従来のセキュリティ対策に加え、AI 固有の観点でのセキュリティ対策も必要となる。

4.2　AI、AI アルゴリズム、AI モデル、AI システム

AI のセキュリティを考えるとき、2.3 節で述べた、AI（機械学習）アルゴリズム、AI（機械学習）モデル、学習用データ、AI モデルが搭載されたシステム（AI システム）が攻撃や対策の対象となる。

システムに関する文脈の中で「AI」と言った場合、「AI アルゴリズム」「AI モデル」「AI システム」の要素を総称する場合、もしくは個別の要素を指す場合がある。

本書では、「**AI システム**」とは、「内部に『AI モデル』を持ち、学習と推論の片方もしくは双方を行うシステム」を指すこととする。「AI モデル」以外の GUI やミドルウェア、OS、インフラなど（サーバ、ネットワーク、IoT 機器など）も含む。AI システムと AI モデルの関係は図 4-1 のとおりである。

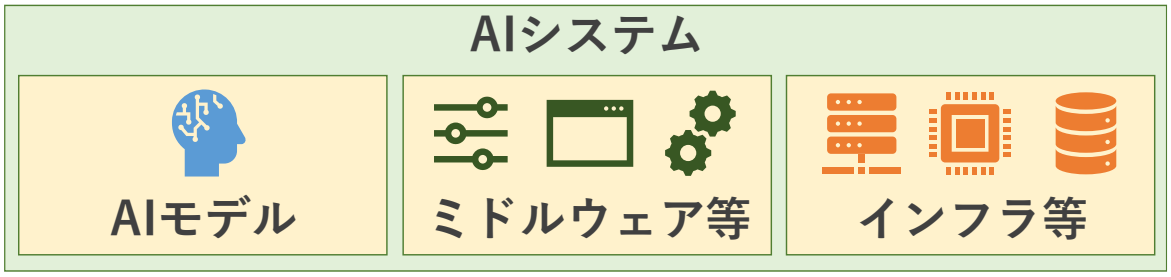

図 4-1 AI システム、AI モデルの関係

---

[38] 例えば、ADVERSA 社による調査<https://adversa.ai/report-secure-and-trusted-ai/>など。AI セキュリティに関する研究論文数が 2015 年では 20 報だったが 2020 年には 1500 報以上となった

[39] <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>

[40] 例えば、Microsoft 社による見解<https://blogs.microsoft.com/on-the-issues/2022/05/03/artificial-intelligence-department-of-defense-cyber-missions/>など。過去 5 年間の間に AI へのサイバー攻撃の増加を観測し、今後もその傾向は続く、との見解

4.3　AI システムにおけるリスク管理

4.1 節でも述べたとおり、将来的には AI システムの普及に伴い AI へのサイバー攻撃のリスクが増大することが予想される。

そのため、AI を含まないシステムと同様に、AI システムに対しても適切なリスク管理を行うことが重要となる。

しかし、AI システムには AI 固有のライフサイクル、資産、脅威が存在する。AI システムのリスク管理を行う際には AI を含まないシステムのリスクに加え、AI 固有のリスクも含めたリスク管理を行う必要がある。

本節では、”AI Cybersecurity Challenges [2]”を参考に、AI 固有のライフサイクル、資産、脅威を紹介し、AI システムにおけるリスク管理の例を説明する。

- ✓　ライフサイクルとの関連を特定する（4.4 節）
- ✓　保護すべき対象の資産を特定する（4.5 節）
- ✓　対象の資産に対する脅威を特定する（4.6 節）
- ✓　脅威に対するリスクを特定し、対策を検討する（4.7 節）

次節以降で、AI システムのライフサイクル、資産、脅威、リスク、対策の概要を説明する。

4.4　AI システムのライフサイクル

2.6 節で述べたとおり、本書では AI ライフサイクルを、次の表のように整理している。

リスク管理の最初のステップは、対象となる AI システムが AI ライフサイクルの各プロセスと、どのように関連しているかを特定することである。

表 4-1 AI ライフサイクル（表 2-1 再掲）

#	プロセス	区分	概要
LC-1	課題設定	企画	AI で解決したい課題を明確にする。課題を解決するための AI の種類を設定したり、求められる AI の精度などを設定したりする。
LC-2	データ収集	開発	AI 構築のために必要なデータを収集する。収集するデータは、組織内データ、公開データ、他組織が提供するデータ、などがある。
LC-3	データ前処理	開発	収集したデータを AI 構築に適した形に変換する。ファイルフォーマットの変換、異常値の除去、ノイズの除去、データの匿名化・仮名化、学習用データの補強（データ拡張）などが含まれる。
LC-4	AI 構築	開発	前処理されたデータをもとに AI を構築する。AI（機械学習）アルゴリズムの選択を行い、学習用データを学習し、AI（機械学習）モデルを構築する。

LC-5	AI 運用	運用	構築された AI を利用する。AI を利用したい環境にデプロイし、AI に推論させたいデータ（推論用データ）を入力して推論させて結果を得る。推論の結果をモニタリングし、時間経過による精度低下などを検知したら、必要に応じて AI のメンテナンスを行う。

4.5 AI システムにおける資産

脅威がもたらされる可能性のある資産および資産のカテゴリーを特定することは、AI システムのリスク管理において非常に重要である。

AI システムにおける資産はデータやソフトウェア、ハードウェア、ネットワークなどの一般的な IT における資産に加え、モデルやプロセッサ、AI アーティファクトなども含む。

“AI Cybersecurity Challenges [2]”を参考に、AI における資産を表 4-2 のカテゴリーに分類し、整理した。カテゴリーごとの資産の具体例を示す。

表 4-2 AI における資産一覧

資産カテゴリー	資産の例
プロセス	データ収集・探索・前処理・拡張、特徴量の選択、モデル構築・学習・チューニング、メンテナンス など
環境・ツール	クラウド、通信ネットワーク、統合開発環境、ライブラリ、AI プラットフォーム など
AI アーティファクト[41]	AI システムアーキテクチャ、データガバナンスポリシー、アクセス制御リスト、ユースケース、ビジネスモデル など
モデル	AI アルゴリズム、ハイパーパラメータ、AI モデル、モデルパラメータ など
アクター・利害関係者	クラウド事業者、データエンジニア、データの所有者や提供者、AI 開発者、データサイエンティスト、エンドユーザ、モデル提供者 など
データ	生データ、ラベル付きデータセット、前処理済みデータ、学習用データ、テストデータ など

[41] 「アーティファクト」は「対象」などと訳される場合もあるが、ここでは「AI アーティファクト」とした。AI に関するポリシー、ルール、ドキュメント、ビジネスケースなど多様なものを含む。表 4-2 はその一例。

4.6 AI システムにおける脅威

脅威を特定することも同じく、AI システムのリスク管理において非常に重要である。

“AI Cybersecurity Challenges [2]”を参考に、AI システムに及ぼされる脅威を表 4-3 のカテゴリーに分類し、整理した（一部重複あり）。カテゴリーごとの脅威の具体例を示す[42] [43]。

表 4-3 AI における脅威一覧

脅威カテゴリー	脅威カテゴリー概要	具体的な脅威例
不正行為	主として AI モデルや学習用データを対象とした窃取、改ざん、破壊などの悪意ある活動	学習用データや AI モデル、機械学習ライブラリなどへのポイズニング（4.7.1 項）、敵対的サンプル（4.7.2 項）の作成や敵対的サンプルによる誤分類や精度の低下、神託攻撃による情報窃取（4.7.3 項）など
意図しない損害	AI システムや AI モデル開発者、AI 利用者の意図しない、資産の破壊、損傷、人への危害などの損害	AI 推論の失敗、ライブラリの設定ミス、データ品質の低下、AI モデルの性能低下など
法的要因	法律や契約に基づく制限、対応義務、不履行に基づく賠償など	運用時のプライバシー侵害、個人情報の流出、法令などで要求されているデータガバナンスポリシーの欠如、SLA 違反など
故障・誤動作	AI システムの一部または全部の故障・誤動作やデータの破損	データや正解ラベルの破損、モデルフレームワークや AI モデルの性能低下など
盗聴・遮断・乗っ取り	主として AI システムのインフラなどを対象とした、盗聴・遮断・乗っ取りなどの活動	データの窃取、推論結果の盗聴、脆弱な暗号化、など
物理攻撃	インフラ、ハードウェア、配線のような物理資産の破壊、無効化、などを目的とする活動	通信網の改ざん、インフラ・システムに対する物理的な攻撃など
機能停止	インフラやプラットフォームなどの予期しないサービスなどの中断または要求する水準を下回る品質	通信網の停止、インフラ・システムの停止など
災害	甚大な損害または人命の喪失を引き起こす突発事故または自然災害	環境現象（気候変動など）、自然災害（地震、洪水、火災など）

[42] AI システムに脅威を及ぼす可能性のある脅威アクターとして、一般的なサイバー攻撃の脅威アクターと同様に、サイバー犯罪者、企業内部関係者、国家主導アクター、テロリスト、ハクティビスト、スクリプトキディ、競合他社などが考えられる

[43] 機械学習システムに対するサイバー攻撃の流れと手法を体系化した MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)も参考になる <https://atlas.mitre.org/>

4.7　AI モデルにおける脅威、脆弱性、対策

本節では、“Securing Machine Learning Algorithms [3]”を参考にして、4.6 節に挙げられている AI システムにおける脅威のうち、「不正行為」、特にその中の AI モデルに関する脅威と脆弱性、対策[44]について概要を説明する。

なお、技術的な詳細や関連研究は、次の Web ページや、資料原書が参考になる。

・参考 Web ページ：AI セキュリティ 情報発信ポータル
<https://www.mbsd.jp/aisec_portal/index.html>
・参考資料原書：“Securing Machine Learning Algorithms [3]”
<https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms>

なお、「盗聴・遮断・乗っ取り」などサイバー攻撃の脅威は通常の IT セキュリティ対策によって、「物理攻撃」「機能停止」「災害」などサイバー攻撃以外の攻撃や自組織のコントロール外の脅威は通常の IT システム向けのリスク管理によって、それぞれ被害を緩和もしくは早期に復旧できる。

また、MLOps（5.7 節）の実践によっても、脅威による被害を緩和もしくは早期に復旧することができると考えられる。

【AI モデルの脅威】

本書では AI モデルへの脅威として、AI モデルへの直接攻撃と学習用データへの攻撃を考えた。サイバー攻撃は 3 種類（ポイズニング（汚染）、回避、神託（情報窃取））に分類した。図 4-2 にそれぞれの攻撃と表 4-1 の AI ライフサイクルとの対応関係を記載している。

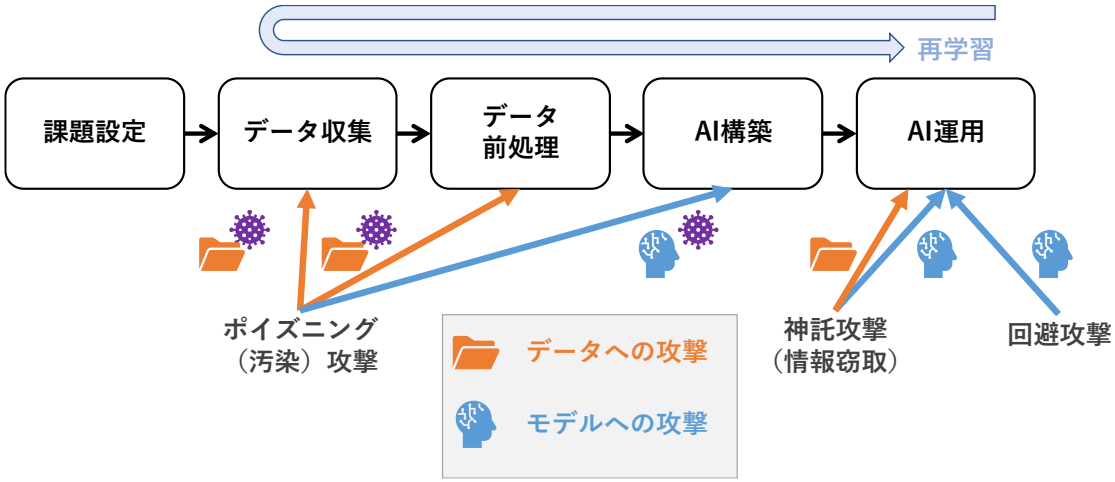

図 4-2 各攻撃と AI ライフサイクルの関係

[44] 本書執筆時点ではドラフトであるが、米国国立標準技術研究所（NIST）の NISTIR 8269 (Draft) A Taxonomy and Terminology of Adversarial Machine Learning <https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf>も参考となる

4.7.1　ポイズニング（汚染）攻撃

【概要】（関連フェーズ：LC-1 課題設定 以外全て）

攻撃者が学習用データや AI モデルに何らかの細工をして、AI モデルの開発者の意図しない推論結果を出力させる攻撃の一種。**本攻撃の対象にはデータと AI モデルの 2 つがある。**

攻撃対象がデータの場合、例えば AI モデルを学習もしくは再学習させる際に、誤った正解ラベルが付けられたデータを攻撃者が学習用データに混入させる。AI モデルがそれを学習してしまうと、再学習後の AI モデルは意図しない推論結果を出力する（図 4-3）。

攻撃対象が AI モデルの場合、例えば、(1) AI モデルに攻撃者が準備した AI アルゴリズムが埋め込まれ、攻撃者の定める特定のフラグ（画像中の文様など）を持ったデータのみ AI モデルが攻撃者の意図する推論結果を出力する場合、(2) AI モデルに使用されている機械学習ライブラリ中のコード実行機能が悪用され、攻撃者によるコード（プログラムやシステムコマンド）が埋め込まれた AI モデルが推論を行う際にそのコードが実行され、データの摂取・改ざんや管理権限の奪取に繋がる場合がある。

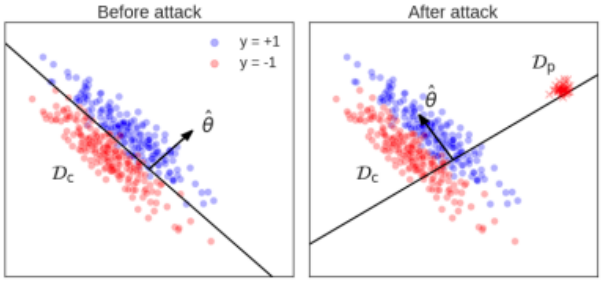

図 4-3 データポイズニングの例[45]

（左）赤い点と青い点が直線（AI モデル）により良く分類されている。（右）ポイズニングされたデータ D_p の存在により、AI モデルの推論結果がおかしくなり、正しく分類されなくなっている。

【脆弱性の例】

- ✓ AI モデルの学習用データの量が不足しており、ポイズニングされたサンプルの影響を受けやすい
- ✓ AI モデル、学習用データ、AI モデルの推論結果などに対して、適切な管理やモニタリングができていないことにより、ポイズニングされていることに気づけない
- ✓ 信頼できない機械学習ライブラリ、AI モデル、学習用データの使用（2.8 節など）

【対応策の例】

- ✓ 学習用データをモニタリングし、疑わしいサンプルを検出して削除する
- ✓ 学習用データの補強（データ拡張）で、ポイズニング攻撃の影響を薄める
- ✓ 学習用データや推論結果をモニタリングし、意図しない振る舞いとなっていないかを確認する
- ✓ 信頼できる機械学習ライブラリ、AI モデル、学習用データを使用する

---

[45] Pang Wei Koh, Stronger Data Poisoning Attacks Break Data Sanitization Defenses, 2018. <https://arxiv.org/pdf/1811.00741.pdf> より引用して抜粋

4.7.2　回避攻撃

【概要】（関連フェーズ：LC-5 AI 運用）

推論用データにノイズが加えられることで AI モデルの推論が誤って行われることがある。加えられたノイズが小さい場合、人間が元のデータとの差異を識別できないことがある。

そうした小さなノイズを見つける攻撃の一種を回避攻撃と呼び、小さなノイズが加えられた推論用データを敵対的サンプルと呼ぶ（図 4-4）。

例えば、製造業における品質管理（2.9.3 項）でこの攻撃が成立すると、良品・不良品判定が適切に行えなくなり、不良品を誤って良品として出荷してしまうなどの被害につながる。

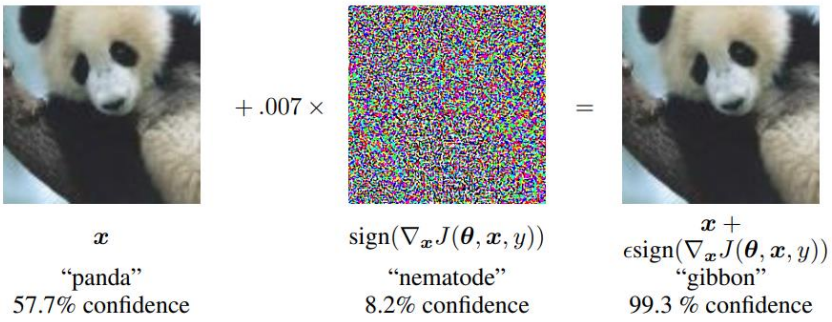

図 4-4 回避攻撃の例[46]

（左）の推論用データは panda（パンダ）に分類されている。小さなノイズ（中央）を加えた（右）の画像を推論用データとして AI に推論をさせると、人間の目には（左）の画像と区別がつかないが、AI は gibbon（テナガザル）と誤って分類（推論）する。（右）の画像を「敵対的サンプル」と呼ぶ。

【脆弱性の例】

- ✓ 敵対的サンプルを検出できない
- ✓ 学習用データに敵対的サンプルを用いる学習の非実施、敵対的サンプルの学習不足
- ✓ 仕様が明らかになっている AI モデル（公開モデルの利用（2.8 節）など）を使用している
- ✓ AI モデルから出力される推論結果以外の情報（信頼スコアなど）が多すぎる

【対応策の例】

- ✓ データが敵対的サンプルであるかどうかを検出するためのツールを導入する
- ✓ 学習用データセットに敵対的サンプルを追加する（敵対的学習）
- ✓ AI モデルの取得経路などを評価し、外部から AI モデル情報を取得されないかを評価する
- ✓ AI モデルから出力される推論結果をユーザーに通知する際、通知する情報は必要最小限にし、ノイズ生成に悪用されるような情報は通知しない

---

[46] Ian J. Goodfellow et al. "Explaining and Harnessing Adversarial Examples," International Conference on Learning Representations （ICLR）, 2015. <https://research.google/pubs/pub43405/>

4.7.3　神託攻撃[47]（情報窃取）

【概要】（関連フェーズ：LC-5 AI 運用）

攻撃者が AI に推論させるデータセットを用意し、そのデータセットとその推論結果を対応させながら観察することで、学習用データや AI モデルに関する情報を窃取する攻撃。学習用データの窃取として、メンバーシップ推論攻撃[48]、インバージョン攻撃[49]（図 4-5）などがある。AI モデルの窃取としてモデル抽出などがある。

学習用データや AI モデルに関する情報を取得するこの攻撃は、その情報を用いて、例えば回避攻撃やポイズニングなど、より有害なタイプの攻撃の前段階となることがある。

図 4-5 学習用データに対する神託攻撃（インバージョン攻撃）の例[50]

インバージョン攻撃では、ランダムな推論用データの入力を繰り返し、AI モデルが推論結果とともに出力する信頼スコアなどの値を観察しながら、ランダムな推論用データを学習用データに近づけていく。
（左）ランダムな推論用データの入力試行により学習用データに近づけられたデータ
（右）元の学習用データ

【脆弱性の例】

- ✓ AI モデルの内部情報（学習により得られたモデルのパラメータなど）が外部から取得可能である（対象が AI モデル）
- ✓ AI モデルから出力される推論結果以外の情報（信頼スコアなど）が多すぎる（対象がデータ、モデルの場合両方）

【対応策の例】

- ✓ 推論用データを AI モデルに入力する前にノイズを加え[51]、攻撃者の観察を妨げる
- ✓ AI モデルから出力される推論結果をユーザーに通知する際、通知する情報は必要最小限にし、学習用データや AI モデル情報の窃取に悪用されるような情報は通知しない

---

47 オラクル攻撃とも呼ばれる

48 <https://www.mbsd.jp/aisec_portal/detail_attack.html#membership_inference_attack>

49 <https://www.mbsd.jp/aisec_portal/detail_attack.html#inversion>

50 Matt Fredrikson, Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, 2015. <https://rist.tech.cornell.edu/papers/mi-ccs.pdf>

51 例えば、「差分プライバシー」におけるノイズ付加など

コラム：AI を使ったサイバー攻撃

AI を使ったサイバー攻撃の例として「ディープフェイク」がある。また、ディープフェイク以外にも既存のサイバー攻撃が AI を活用することでより高度に行われるとの研究[52]もある。

AI を使ったサイバー攻撃を、第 3 章で紹介したような AI を活用したサイバーセキュリティ対策で防御するなど、サイバー攻撃の高度化・防御手法の高度化の応酬が続いている。

本コラムでは AI を使ったサイバー攻撃の例として「ディープフェイク」について簡単に例を説明する。攻撃者側も AI をサイバー攻撃の武器として活用しており、AI の活用を含めたより一層の防御能力の向上が必要である。

➢ ディープフェイク

「ディープフェイク」は、ディープラーニング技術を応用したメディア合成技術である。従来の合成技術では難しかった、動画や音声の精度の高い合成を行えるのが特徴である。

「実際には行っていない動作」や「実際にはしていない発言」の動画や音声を得られるうえ、付帯情報なしに人間がそれらを偽物と見破ることが困難であるため、様々な目的のため悪用されている。

例えば、有名人の偽の動画によって尊厳を傷つける、政治家の発言をねつ造して世論を操作しようとする、偽の CEO の声によって担当者をだまし不正に送金させるといった事例がある。

ディープフェイクの対策を技術的な方法によって行うことは難しい。わずかに残る不自然な点を利用して見破る手法も研究されているが、結局はいたちごっこであり本質的な対策とは言いがたい。ディープフェイク作成者が学習用データを入手できないよう、画像や音声を非公開にするのは効果的だが、有名人の場合には現実的ではない。

Google による、既知のディープフェイク生成方法により作られた画像[53]
ディープフェイク対策の研究ためのデータセットとして公開されている

---

[52] 例えば、Yisroel Mirsky et al. "The Threat of Offensive AI to Organizations", 2021 <https://arxiv.org/pdf/2106.15764.pdf> 「AI による bot 操作」などが挙げられている

[53] Contributing Data to Deepfake Detection Research <https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html>

コラム：AI で AI を守る、AI の検知を回避する

第 3 章、第 4 章でそれぞれ説明したとおり、AI を活用したサイバーセキュリティ対策と AI のセキュリティは異なる観点の議論である。しかし、両者は無関係とは限らない。以下に示す例においては、両者は密接に関係している。

➢ AI、AI システムを守るための AI 活用セキュリティ製品の利用

第 3 章で説明した AI 活用セキュリティ製品で守る対象のシステムが、第 4 章で説明した AI システムである、という場合、両者には守る・守られるという関係がある。AI 活用セキュリティ製品で守るシステムがたまたま AI システムであるという見方もできるが、AI システムへアップロードされるデータやファイルを AI で検査することで、異常なデータの入力や異常な振る舞いの検知ができ、結果的に AI システムのセキュリティが担保される場合もあると考えられる。

➢ AI 活用セキュリティ製品の検知回避

セキュリティ製品に検知されることを回避するために、マルウェアがセキュリティ製品を停止したり、もともと環境に存在する正規のツールを悪用したりするなどの動作を行う場合がある。これに加えて、マルウェアが AI 活用セキュリティ製品の AI による検知ロジックを回避することが可能であるという事例も報告されている[54]。

AI 活用セキュリティ製品を開発しているベンダーも当然、こうした動作に対する追加策を日々研究・実装しているが、結局はいたちごっこなってしまう。

AI の限界を理解し、性能を過信することなく、多層防御で組織の資産を守っていく必要がある。

➢ 顔認証システムへの回避

重要インフラ等においては、物理セキュリティとして、特定の物理区画に入場する際に、顔認証を求める場合がある。近年では顔認証を AI で行うというサービスが出てきているが、AI の顔認証をごまかす手法も存在する。例えば「敵対的メガネ[55]」と呼ばれるもので、特殊なデザインをしたメガネをかけることで AI の顔認証をごまかし、別人になりすまして入場することができる。

こうしたリスクを認識した上で、必要に応じて人間による確認を追加するなど、AI の限界を理解した上で AI を上手く活用し、組織の資産を守っていく必要がある。

---

[54] 例えば、<https://jvn.jp/vu/JVNVU98738756/>

[55] NRI セキュアテクノロジーズ ブログ<https://www.nri-secure.co.jp/blog/hostile-sample>

コラム：OT 分野における AI 活用とセキュリティ

一般的に ICT（Information and Communication Technology；情報通信技術）分野と対比して、鉄鋼・化学などの製造、電力・ガス・水道・交通などの社会インフラを支える制御技術を OT（Operational Technology：運用技術）と呼ぶ。OT 分野におけるシステム（OT システム）では、システムの停止により国民生活に大きな影響を及ぼしうる性質上、情報セキュリティの 3 要素[56]の中で特に可用性の側面が重要視される。ただし単に稼働し続けていれば良いという話ではなく、例えば電力インフラの電圧や周波数、水道インフラの水質などの適正な範囲は法令で定められており、また製造業でも良品として出荷できる基準が存在するなど、情報セキュリティにおける「完全性」ほど厳格ではないものの「然るべき品質が保たれている」ことを前提とした「可用性」であることに注意が必要である。

このような背景から、従来のシステムと比較してどのようなメカニズムで結果が出力されたのか分かりにくい AI のテクノロジーを OT システムに取り入れることは、原因の特定が困難な（=復旧が困難かつ説明責任も果たせない）可用性の喪失を引き起こすおそれがあり、ICT 分野に比べて AI の導入に関する議論は進んでいなかった。しかし、近年になって OT 分野への AI の導入が現実的に検討され始めている。背景には次のような業界を取り巻く事情が影響していると考えられる。

- ✓ 人材の流動化が進む中で、熟練者の技能に依存した事業運営のリスクが高まっていること
- ✓ 団塊の世代の大量退職や労働生産人口の減少に伴い技術継承が難しくなっていること
- ✓ 自由化や国際競争力確保のためのコスト削減に対する継続的な要請があること
- ✓ ICT 分野での AI 技術の利用拡大により、説明可能性の確保を含む AI システム利用のノウハウが蓄積されてきたこと[57]

➢ **OT 分野での AI 利用の例**

OT 分野での AI 活用事例としては第 2 章の 2.9.3 項「品質管理」、2.9.4 項「製造プロセスの置き換え」、2.9.5 項「機器の障害・寿命予測」のほかに、今後期待される用途として、複雑な系の制御が挙げられる。

例として、電力インフラの供給エリアの中に散在する各地の電圧や流れる電力の量の需給バランスのパラメータに対して、それらを調整する多数の手段が同じくエリア内に散在する状況を考える。これまでは熟練のオペレーターがどの手段が全体最適かを、知識や過去の経験をもとに判断して必要な制御を行っていたが、このオペレーターの判断を教師として学習させた AI でこれらの業務を補助または代替することが考えられる。また、業界の自由化に伴う新電力などのプレーヤーの増加や自然変動

[56] 機密性、完全性、可用性の 3 つを指す。以下のコラム部分が参考となる
<https://www.soumu.go.jp/main_sosiki/cybersecurity/kokumin/intro/intro_security.html>

[57] 例えば、三菱電機「制御の根拠を明示できる AI 技術」を開発
<https://www.mitsubishielectric.co.jp/news/2021/pdf/1214.pdf>

再生可能エネルギーの大量導入、VPP（仮想発電所）[58]といった新技術の導入により、高度かつ複雑化する次世代電力システムの最適制御を実現する技術としても注目されている[59]。

また別の例として、道路インフラにおいて、車両から収集される位置や速度などのセンサー情報に基づき、AI による渋滞予測により道路信号を最適制御する、といった試みも本書執筆時点で実証段階まで進んでいる[60]。

なお、これらの複雑な系を取り扱う方法としては物理計算に基づくシミュレーションが一般的だが、処理するモデルによっては AI を活用してより高速短時間で結果を予測できるとする研究が行われており、計算の高速化と計算資源の有効活用の面でも期待されている[58] [59]。また、シミュレーションで生成される膨大なデータを AI で処理し、望ましい結果を得るための条件を推定させる、といったシミュレーションと AI を相互に連携させる研究も行われている[61]。

- **OT 分野での障害の影響とリカバリープラン**

OT 分野に導入された AI システムで障害が発生した場合、ICT 分野と同様のシステム復旧に向けた試みに加えて、冒頭に記載した「可用性」が重視される OT 分野の事情を加味する必要がある。

プラントの停止によるサプライチェーンへの影響や重要インフラを利用する国民生活への影響を最小限に抑えるため、可能であれば AI 技術を使用していない自動制御システムへの切り替えによる縮退運用[62]、または人間の手動運用への切り替えがリカバリープランとして考えられる。

また、システム停止に至らないものの AI の出力に明らかな異常が見られるような場合、「可用性確保のために運用を継続するか」「品質が許容できないためシステムを停止（代替）するか」の非常に難しい判断を迫られる状況も考えられるため、事前に停止の基準と責任者（判断者）を整理し取り決めておくことが望ましい。

---

[58] Virtual Power Plant の略称。地域に点在する小規模な再エネ発電や蓄電池、燃料電池などの分散された発電設備やシステムを、IoT 技術を活用してコントロールし、あたかも 1 つの発電所のように機能させる技術

[59] 経済産業省・資源エネルギー庁：「電力ネットワークの次世代化について」<https://www.meti.go.jp/shingikai/enecho/denryoku_gas/denryoku_gas/pdf/048_05_01.pdf>

[60] マイナビ TECH+：「NEDO、人工知能を活用した信号制御システムの実証実験を実施と公表」<https://news.mynavi.jp/techplus/article/20220425-2329503/>

[61] 新エネルギー・産業技術総合開発機構：「計算シミュレーションと AI を連携させ、仮想実験環境を構築」<https://www.nedo.go.jp/news/press/AA5_101424.html>

[62] 通常使用する方式や系統が正常に機能しなくなったときに、機能や性能を制限したり別の方式や系統に切り替えたりするなどして、限定的ながら使用可能な状態を維持すること

さらに、実際のシステム停止による影響の大小や、代替手段による縮退運用の可否については、事前に AI システムの障害を想定した全体システム設計がなされているかといった要素に加えて、AI が代替する業務の性質も影響する。

例えば、「品質管理」、「機器の障害・寿命予測」の例では、それぞれ「人の目視精度を超えた品質判定」、「機器のコンディションの実態に即したリアルタイムの保守管理」がそれぞれ AI によって実現されており、AI の能力を活用した付加価値が生み出されているといえる。

この場合、AI システムの停止により、AI 導入で得られていた価値（人的コストや保守コストの最適化）が一時的に失われてしまうことになる。しかし、「品質管理」の例では「従来どおり人が目視で判断する」または「単純なしきい値による自動判定」、「機器の障害、寿命予測」の例では「従来どおり定期的な点検や平均耐用年数による予防保全」に運用を縮退することで、ある程度の業務継続が期待できる。

- **業務の置き換えとしての AI の導入**

一方、「品質管理」、「製造プロセスの置き換え」、前項「複雑な系の制御」のような活用例では、個人の高度なノウハウに依存（属人化）していた業務を置き換えることを目的とした導入であることから、AI システムへの移行により、熟練の人材を確保しておく必要性がなくなり、場合によっては退職や転職、技術継承の機会消失によって社内に AI に移行していた業務を代行できる人材がいなくなってしまっている事態も想定される。こうした場合には業務の人間系での代替ができず、システム障害が企業の事業継続により深刻な影響を与える可能性がある。

こうした自動化に伴う技術力低下・喪失の議論は従来の自動制御の導入時から行われていたが、AI はより高度なノウハウの置き換えのための利活用が期待されることから、より深刻な影響が生じる可能性がある。このため、BCP[63]に直結する業務の置き換えとして AI の導入を検討する際には、AI システムの停止の可能性を想定して、コストをかけてでも障害発生にも対応できる強固なシステムを構築するのか、停止した場合に業務を代替できる人材をある程度社内に残すのか、残す場合、どのように技術レベルの維持と継承を行っていくのか、といった要素を十分に検討する必要がある。

これは言い換えると、BCP に直結する業務において AI をどのように位置付け、活用するかという課題に対する、組織としての明確なビジョンと合意形成の重要性を意味する。

[63] Business Continuity Plan の略称。事業継続計画と訳される。組織が危機的な状況に置かれた場合であっても、重要な業務を継続できるようにしておくための計画

第5章　　関連事項

5.1　AI 社会原則

AI の活用は社会に多大な利益をもたらすが、その一方で、社会に対する影響力も大きい。そのため、AI の社会への適切な実装が求められる。そのような AI の適切な社会実装について方針を与えるものが「AI 社会原則」である。

内閣府 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 [5]」では、基本理念から始まり、ビジョン（AI-Ready な社会）、人間中心の AI 社会原則が提示されている（図 5-1（左））。

基本理念として実現を追求すべき社会として①「人間の尊厳が尊重される社会」、②「多様な背景を持つ人々が多様な幸せを追求できる社会」、③「持続性ある社会」、が挙げられている。

また、ビジョン「AI-Ready な社会」とは、「社会全体が AI による便益を最大限に享受するために必要な変革が行われ、AI の恩恵を享受している、または、必要な時に直ちに AI を導入しその恩恵を得られる状態にある、『AI 活用に対応した社会』」を意味すると説明されている。

人間中心の AI 社会原則として①「人間中心の原則」、②「教育・リテラシーの原則」、③「プライバシー確保の原則」、④「セキュリティ確保の原則」、⑤「公正競争確保の原則」、⑥「公平性、説明責任及び透明性の原則」、⑦「イノベーションの原則」、が挙げられている（図 5-1（右））。

第 3 章の留意点の検討にあたっては、この「人間中心の AI 社会原則」の各原則も参考としている。

本書ではそれらの原則の詳細には立ち入らないが、大いに参考になる。

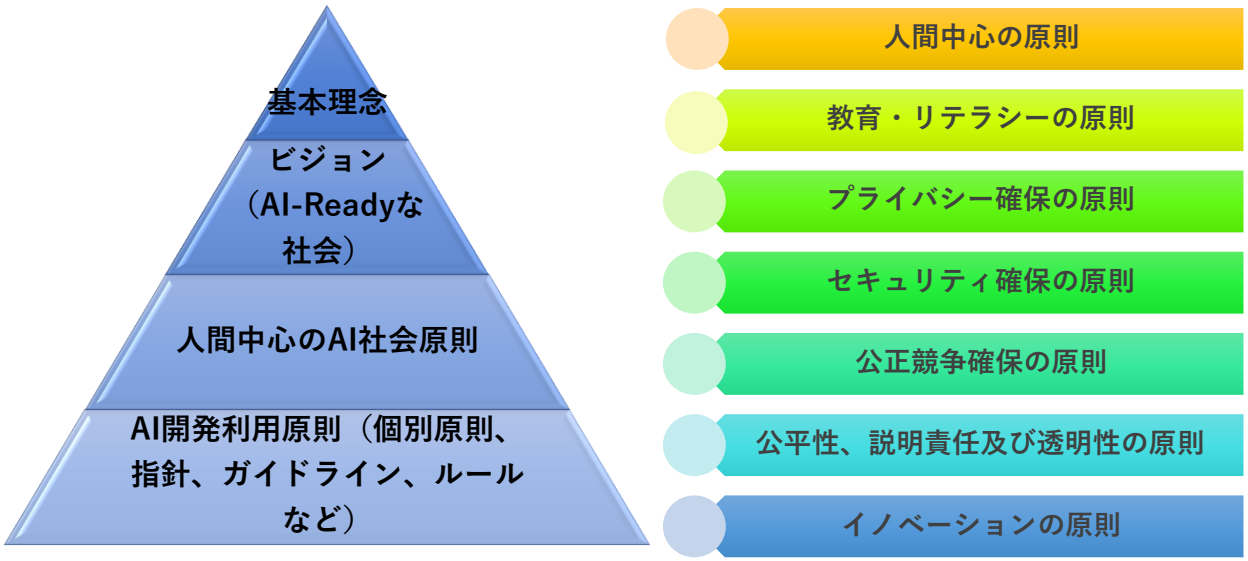

図 5-1 人間中心の AI 社会原則の階層（左）と各原則（右）
「人間中心の AI 社会原則」をもとに作成

5.2　AI ガバナンス

経済産業省 AI 原則の実践の在り方に関する検討会「AI 原則実践のためのガバナンス・ガイドライン ver. 1.1 [6]」には、

AI 社会原則は、①人間中心の原則、②教育・リテラシーの原則、③プライバシー確保の原則、④セキュリティ確保の原則、⑤公正競争確保の原則、⑥公平性、説明責任及び透明性の原則、⑦イノベーションの原則の 7 つの原則から構成される。この AI 原則実践のための企業ガバナンス・ガイドライン（略して「AI ガバナンス・ガイドライン」という。）

との記述がある。このことから、「AI ガバナンス」とは、「AI 社会原則を実践するための企業（組織）ガバナンス」と解釈することができる。人間中心の AI 社会原則はあくまで原則であり、組織がその原則に則った活動するためには、各組織に適した管理体制の構築、すなわち AI ガバナンスが必要である。

「AI 原則実践のためのガバナンス・ガイドライン ver. 1.1」では、AI ガバナンスに関する行動目標と、参考情報としての仮想的な実践例が豊富に掲載されている。

例えば、その中の行動目標 1 － 1 として「AI システムがもたらしうる正負のインパクトを理解する」というものが挙げられている。多くの情報源で、正のインパクト（AI のメリットや価値など）については情報が提示されている。この行動目標は、負のインパクトやリスク[64]も忘れてはいけないことを示している。これはセキュリティと AI を考えるうえでも重要なことである。

第 3 章の留意点の検討にあたっては、この行動目標のほか、その他の行動目標も参考としている。本書ではそれらの行動目標の詳細には立ち入らないが、非常に参考になる。「AI 原則実践のためのガバナンス・ガイドライン ver. 1.1」は仮想的な実践例も詳細かつ豊富であるため、そちらも参考になる。

また、先進的な企業では、AI ガバナンスに関し、意欲的・積極的に取り組んでいる[65]。AI ネットワーク社会推進会議「AI ガバナンスに関する取組事例 [18]」において、取り組みを行っている企業と、その取り組みが紹介されている。

AI ガバナンスに関する取組の例として、セキュリティに関する取り組みも紹介されている。AI に対する攻撃を検知する技術や攻撃を防御する技術についての研究開発や、運用時に敵対的攻撃から防御する技術（4.7 節で述べた内容の一部）などが挙げられている。

---

64 本書執筆時点ではドラフトであるが、AI リスクマネジメントのフレームワークとして米国国立標準技術研究所（NIST）の AI Risk Management Framework < https://www.nist.gov/itl/ai-risk-management-framework>も参考となる。

65 例えば、NTT データ <https://www.nttdata.com/jp/ja/services/ai/governance/>

5.3 AI 倫理

AI は、その処理がブラックボックスであり、かつ、自動で利用・運用されるケースが多い。そのため、AI が出力した結果が人間の想定外の結果、意図しない結果となる可能性があり、問題となる場合がある。

例えば、AI が出力した結果が、人種差別、性差別的なものとなっている場合がある。例としては 4.1 節で述べたチャットボットの差別的発言などがある。4.7 節で述べたセキュリティ対策により、不正な入力（ポイズニング）などを防ぐとともに、学習用データに人種に関する情報を入れない、などセキュリティ以外の観点の対応が必要となる。

また他にも、個人を不当に取り扱ってしまう場合がある。例えば、内部不正検知において、疑わしい個人が特定されて通知される場合、事実確認などの検証を行わずにその個人を罰する、などがある。

G20 AI 原則 [4]、人間中心の AI 社会原則 [5]、AI 原則実践のためのガバナンス・ガイドライン [6] においても AI 倫理について言及されている。

また、AI 倫理に関するガイドラインを定め、運用している企業[66]もある。今後もそうした企業が増えていくこと、もしくは読者の組織においても対応が求められることが想定される。

5.4 説明可能 AI

ブラックボックスである AI（2.3 節）においては、その出力（推論結果）の根拠を説明することが難しい場合がある。一方、AI の出力に対して根拠を求められる場合がある。例えば、医療分野において診断を AI で行う場合、患者へ診断結果を説明しなければならない場合など、AI の出力とともにもその根拠も合わせて求められる場合がある。

近年、結果を説明ができる AI の開発が進んでいる。そうした AI は説明可能 AI[67]（Explainable AI; XAI）と呼ばれている。入力値の変更に対する推論結果の変化から重要な要素を推定する、開発段階から説明が可能な AI アルゴリズムを採用する、などの手法がある[68] [69]。

こうした説明可能 AI が普及することで、これまで推論結果しか活用できなかった事例において、さらに説明の根拠を活用してメリットを享受できる場合がある。

例えば、機器の障害・寿命予測（2.9.5 項）などでは、使用する AI が説明可能 AI である場合、故障時期の予測に加え、その根拠（例えば、故障につながる機器の動作など）がわかることで機器の改善につなげることができる。

また、AI 活用セキュリティ製品においても説明可能 AI が採用されることで、AI の出力が解釈しやすくなり、SOC 業務の効率化につながるなどが考えられる。

---

[66] 例えば、富士通 <https://www.fujitsu.com/jp/about/research/technology/aiethics/> や、日立製作所 <https://www.hitachihyoron.com/jp/archive/2020s/2021/sp/sp01/index.html>など

[67] 解釈可能 AI などとも呼ばれる場合もある

[68] 人工知能学会 <https://www.ai-gakkai.or.jp/resource/my-bookmark/my-bookmark_vol33-no3/>

[69] 米国防高等研究計画局（Defense Advanced Research Projects Agency; DARPA）による 3 つのアプローチもよく引用されている。
< https://www.darpa.mil/attachments/XAIIndustryDay_Final.pptx> B.1 Explainable Models など

5.5 AI の品質

AI システムは、従来の AI を搭載していないシステムとは異なり、AI の出力（推論結果）の根拠を説明できない、学習用データと推論結果の対応付けが難しく想定外の結果が出力される場合がある、運用の仕方によっては AI モデルが随時更新されていく（2.3 節）、などの特徴がある。

AI の利用の増加や、AI を活用したサービスの市場展開が進むに伴い、AI の有効性や信頼性などの品質を定量的に評価したいというニーズが高まってきている。

そうしたニーズに対して、AI（機械学習）の品質に関するガイドラインが公開されている。例えば、産業技術総合研究所から「機械学習品質マネジメントガイドライン 第 2 版［15］」が公開されている。「機械学習品質マネジメントガイドライン 第 2 版」では、

機械学習システムにおける品質そのものを、

- *システムがその全体として利用時に満たすことが期待される「利用時品質」*
- *システムのうち機械学習で構築された構成要素が満たすことが期待される「外部品質」*
- *機械学習による構成要素が固有に持つ「内部品質」*

の 3 つに分けて理解し、機械学習要素の「内部品質」の向上を通じてその「外部品質」を必要となるレベルで達成し、最終的なシステムの「利用時品質」を実現するものと整理する

としている。AI の品質を議論する際にこの分類は有用である。次に示すガイドラインなどでも言及されている。

また、大学や大手 IT ベンダーなどに所属するメンバーから構成される、AI プロダクト品質保証コンソーシアム（略称：QA4AI コンソーシアム）の「AI プロダクト品質保証ガイドライン 2021.09 版［19］」が公開されている。

分野に特化したガイドラインとしては、石油コンビナート等災害防止 3 省連絡会議（経済産業省、総務省消防庁、厚生労働省）の「プラント保安分野 AI 信頼性評価ガイドライン 第 2 版［20］」が公開されている。今後も様々な分野でこうした分野特化型のガイドラインが出てくることが考えられる。

契約の観点では日本ディープラーニング協会「契約締結における AI 品質ハンドブック［21］」が公開されている。

これらのガイドラインは今後も新たに公開されたり、更新されたりする場合があるため、継続した情報収集が必要である。

また、AI 品質に関するガイドラインを定め、運用している企業[70]もある。今後もそうした企業が増えていくこと、もしくは読者の組織においても対応が求められることが想定される。

さらに、AI の品質を保つことは、結果として、敵対的サンプル（4.7.2 項）に対する耐性向上につながる場合もあり、セキュリティの観点でも重要である。積極的に敵対的サンプルを学習用データとして使用することで、AI の堅牢性を向上させる手法（敵対的学習）（4.7.2 項）も存在する。

[70] 例えば、日本電気（NEC）< https://jpn.nec.com/press/201912/20191210_02.html>

5.6　プライバシー

5.1 節の AI 社会原則でも挙げられているとおり、AI の活用においてプライバシーの確保は重要である。また、AI の活用の事例としても、2.9.1 項で紹介したスマートスピーカーや、スマートウォッチ、室内センサーなどの IoT デバイスから収集されたデータによる健康管理など、パーソナルデータを学習用データ、もしくは推論用データとして利用する活用事例も多い。

パーソナルデータには、個人情報保護法や欧州の GDPR（General Data Protection Regulation; 一般データ保護規則）などの法令などで守るべき対象として定められているデータのほか、個人に関するより一般的なデータ（個人の位置情報など）が含まれている。

パーソナルデータを AI で利用する際には、不要なパーソナルデータを収集・保存しないこと、データを適切に管理（保存・通信の暗号化、適切な保管期間の設定、廃棄など）することが必要である。また、AI が出力した結果からの個人に関する情報の特定（メンバーシップ推論（4.7.3 項）など）が行われないように注意が必要である。特に AI の場合には、関連する多種類のパーソナルデータを処理する場合も多く、各種のデータの結合により、意図せず個人の特定につながってしまう可能性も高くなる。

プライバシーを保護するための技術はいくつか存在する[71]が、ここでは、データに対する技術である「匿名化」「仮名化」を紹介する。

「匿名化（Anonymization）」は、AI で処理したいデータを、個人が識別できないように加工する処理である。一方「仮名化（Pseudonymization）」は「ほかの情報と照合しない限り」個人が識別できないように加工する処理である。両者は異なる概念であるが、区別されずに使用されている場合も散見される。また、加工の方法を誤ると法令違反となる場合があるため注意が必要である。加工の方法についてはガイドライン [22]や法令・ガイドラインに準拠した加工を提供するソリューション[72]が存在するため、それらを活用することも有効であると考えられる。

本節では AI の関連事項としてプライバシーについて述べたが、厳密な用語の定義や法令上の取り扱いは行っていない。必要に応じて、法令やプライバシーに関するガイドライン [23]などを参照いただきたい。

---

[71] 例えば、日本総合研究所 「プライバシー強化技術の概説と動向」 に記載されている技術など <https://www.jri.co.jp/MediaLibrary/file/column/opinion/pdf/13005.pdf>

[72] 例えば、NTT テクノクロス<https://www.ntt-tx.co.jp/products/anontool/>や NEC ソリューションイノベータ< https://www.nec-solutioninnovators.co.jp/sl/danony/>など

5.7 MLOps[73]

MLOps とは、組織により定義は異なるが、おおむね「AI（機械学習）の開発チームと運用チームが互いに協働して、AI（機械学習）モデルの開発から運用までのプロセス、すなわち AI ライフサイクルを自動化により円滑に進めるための体制や技術、考え方」などのようにまとめられる。例として、ある企業における MLOps 取り組みを図 5-2 に示す。

類似の用語に DevOps、DevSecOps があり、大まかには、それらの AI（機械学習）版であると言える。

上で述べたとおり、MLOps の定義は組織で異なるが、例えば、IT エンジニア向けサイト「@IT[74]」では、MLOps の要素として次の要素が挙げられている。

- ✓ モデルの再利用とバージョニング
- ✓ モデルの挙動検証（単体テストなど）とパフォーマンス検証
- ✓ モデルのデリバリー／デプロイ
- ✓ モデル実行のログ出力と分析
- ✓ モデル実行のモニタリング

モデルのパフォーマンス検証やモニタリングは 4.7.1 項で述べたポイズニング（汚染）攻撃への対策となりうる。また、モデルのバージョニングが行われており、モデルがすぐにデリバリー、デプロイできることなどは、AI システムが攻撃された場合に、迅速な復旧につながる。

MLOps は AI（機械学習）システムの効率的な運用はもちろん、セキュリティ面へも寄与するため、積極的に取り組むことが望ましい。

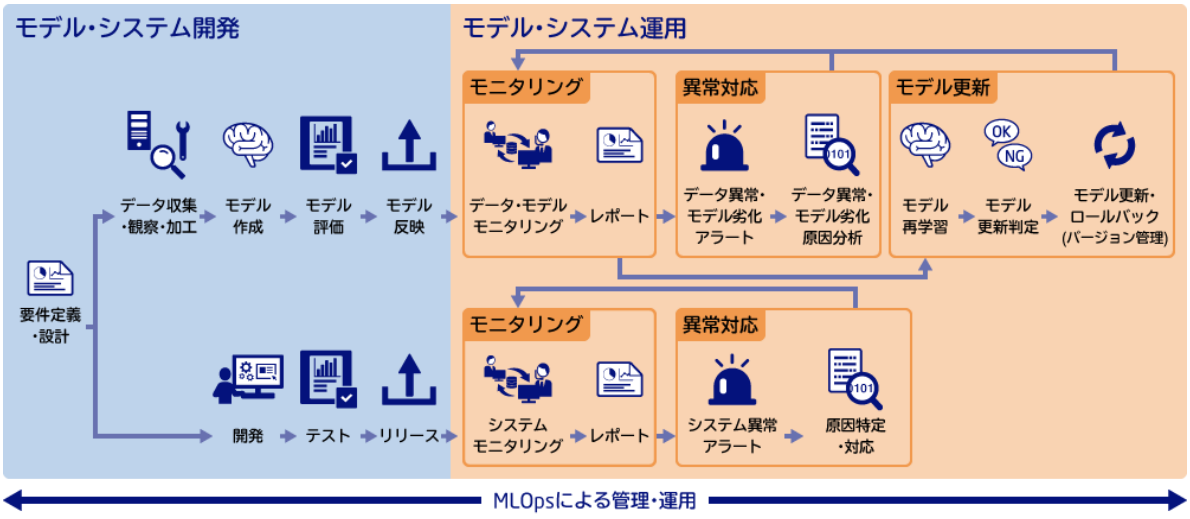

図 5-2MLOps の例[75]

---

[73] 似た用語に「AIOps」という用語がある。「AIOps」は AI により IT システムなどの運用効率化などを図ることを指すものであり、AI のライフサイクルに関する用語ではないことに注意。

[74] <https://atmarkit.itmedia.co.jp/ait/articles/1911/21/news018.html>

[75] 日本電気（NEC） 「NEC MLOps サービス」<https://jpn.nec.com/mlops/index.html>

第6章　　おわりに

6.1　まとめ

本書では、まず第 2 章で、セキュリティと AI を考えるために必要な AI（機械学習）知識を説明した。

第 3 章では、「AI 活用セキュリティ製品」を組織で企画・導入・運用する上での留意点を説明した。これらの留意点を参考に検討を進めていただくことで、AI 活用セキュリティ製品に関する課題が事前に把握・解決され、円滑に組織で活用されることを期待する。

第 4 章では、AI のセキュリティについて説明した。AI システムの脅威、資産、AI モデルに対する脅威、脆弱性、対策を説明した。これらは網羅性を保証しているものではないが、AI のセキュリティに関する気付きとしていただき、組織の AI のセキュリティ向上に役立てていただければ幸いである。

第 5 章では、セキュリティと AI に関連する事項を説明した。どれもセキュリティとの関連が強く、今後これらの検討に関与するセキュリティ関係者も増えてくると考えられる。そうした方々の参考になれば幸いである。

全体を通して、本書を読む前と比べて、少しでもセキュリティと AI について詳しくなっていただき、本書の目的である、読者の組織のセキュリティ分野における適切な AI の活用と、組織で活用される AI のセキュリティ向上を推進が達成されれば幸いである。

6.2　本書の課題と制約

6.2.1　対策ができない脅威、脆弱性および残存リスク

本書で紹介した AI への脅威、脆弱性の中には、現時点における技術、または経済効率・運用の観点で、直接的な対策が難しいものも含まれている。ただし、それらを残存リスクとして認識し、リカバリープランの策定やインシデント発生時の対応方法を事前に検討しておくことで損害を最小限にすることはできる。

繰り返しになるが、本書の目的は、各組織でのセキュリティ分野での AI 活用と AI のセキュリティリスクに対する対策を推進することにあり、対策の難しい脅威、脆弱性を示して AI の活用にブレーキをかけるものではない。

AI の技術は日進月歩であり、攻撃に対してより耐性のある AI アルゴリズムや対策も現れることが十分に予想される。残存リスクを認識することにより、それらを自組織に適用するためのきっかけとしていただき、よりセキュアな AI 活用が達成されることを望んでいる。

6.2.2　各組織での活用

本書ではセキュリティと AI について、特定の組織や個別の製品、個別の AI システムについて言及しているわけではないため、状況に応じて適切に読み替えをしてほしい。また、技術的な詳細にも踏み込んでいない。更に興味のある読者には、各章で紹介した文献などが参考となる。

6.3 謝辞

本書の作成にあたり、独立行政法人 情報処理推進機構 産業サイバーセキュリティセンター 中核人材育成プログラム講師の、満永 拓邦先生、門林 雄基先生、宮本 大輔先生、松田 亘先生、藤本 万里子先生には、本書の元となるプロジェクトのメンター・講師として、ご指導・ご助言、ご支援を賜りました。改めて御礼申し上げます。

また、有識者として、三井物産セキュアディレクション株式会社の高江洲 勲様にも有益なご助言および査読をいただきました。改めて御礼申し上げます。

なお、中核人材育成プログラム 3 期生の Security for AI に関する卒業プロジェクトのメンバーからもご助言および検討成果のご提供をいただきました。この場を借りて御礼申し上げます。

プロジェクトメンバー

本書は、独立行政法人 情報処理推進機構 産業サイバーセキュリティセンター 中核人材育成プログラムにおける卒業プロジェクト「セキュリティ関係者のための AI ハンドブック」の成果物として作成されました。

＜プロジェクトメンバー＞

（◎はリーダー、○はサブリーダー）

伊藤 伸也
浮田 裕基
浮本 敦
○内野 隆志
澤田 裕介
○篠原 隆
◎鈴木 真徳
○田原 淳平
野上 晋平
行 良治

参考文献

[1] 総務省, 「令和元年版 情報通信白書」, 2019.

[2] ENISA, “Artificial Intelligence Cybersecurity Challenges”, 2020.

[3] ENISA, “Securing Machine Learning Algorithms”, 2021.

[4] G20, 「AI 原則」, 2019.

[5] 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」, 2019.

[6] AI 原則の実践の在り方に関する検討会,「AI 原則実践のためのガバナンス・ガイドライン ver. 1.1」, 2022.

[7] National Cyber Security Centre, “Intelligent security tools”, 2019. [オンライン].

[8] 情報処理推進機構, 「DX 白書 2021」, 2021.

[9] 情報処理推進機構, 「AI 白書 2020」, 2020.

[10] AI 白書編集委員会, 「AI 白書 2022」, 2022.

[11] NIST, “NIST Special Publication 1270 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence”, 2022.

[12] Center for Security and Emerging Technology(CSET), “Key Concepts in AI Safety: Robustness and Adversarial Examples”, 2022.

[13] 科学技術振興機構, 「人工知能研究の新潮流 ～日本の勝ち筋～」, 2021.

[14] AI ネットワーク社会推進会議, 「報告書 2021 ～ 『安心・安全で信頼性のある AI の社会実装』の推進 ～」, 2021.

[15] 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」, 2021.

[16] AI ネットワーク社会推進会議, 「AI 利活用ガイドライン」, 2019.

[17] 経済産業省, 「AI・データの利用に関する契約ガイドライン AI 編」, 2018.

[18] AI ネットワーク社会推進会議, 「AI ガバナンスに関する取組事例」, 2021.

[19] AI プロダクト品質保証コンソーシアム,「AI プロダクト品質保証ガイドライン 2021.09 版」, 2021.

[20] 石油コンビナート等災害防止 3 省連絡会議（経済産業省、総務省消防庁、厚生労働省）, 「プラント保安分野 AI 信頼性評価ガイドライン 第 2 版」, 2021.

[21] 日本ディープラーニング協会, 「契約締結における AI 品質ハンドブック」, 2021.

[22] 個人情報保護委員会, 「個人情報の保護に関する法律についてのガイドライン」, 2021.

[23] 総務省・経済産業省, 「DX 時代における企業のプライバシーガバナンスモデルガイドブック ver1.2」, 2022.

[24] 内閣サイバーセキュリティセンター（NISC）, 「サイバーセキュリティ 2021」, 2021.

[25] AI 原則の実践の在り方に関する検討会, 「我が国の AI ガバナンスの在り方 ver. 1.1」, 2021.

用語集

用語	意味・解説
脅威	システムや組織に損害を与える可能性がある潜在的な原因。「物理的脅威」「技術的脅威」「人的脅威」などに分類できる。
脆弱性	「脅威」が付け入ることができる弱点。ソフトウェアのバグやセキュリティホールが該当する。
リスク	システムや組織などが損害を受ける可能性。情報漏洩やサイバー攻撃が例として挙げられる。
パーソナルデータ	個人に関するあらゆる情報を指す。個人情報保護法で定義される個人情報だけでなく位置情報、購買情報、スマートウォッチのログなどのデータも含まれる。
DevOps	Development（開発）と Operations（運用）を合わせた用語。開発チームと運用チームがソフトウェアライフサイクル全体に渡って連携し、ソフトウェアの迅速かつ高頻度なリリースを可能にするための文化や組織体制の構築、プロセス、手法を指す。
DevSecOps	DevOps に対して Security も加えた用語。DevOps にセキュリティチームも参画し、ソフトウェアライフサイクル全体のセキュリティを確保しながら、ソフトウェアの迅速かつ高頻度なリリースを可能にするための文化や組織体制の構築、プロセス、手法を指す。
SOC	Security Operation Center の略称。セキュリティ製品などのアラートをもとにサイバー攻撃の検知や分析を行い、その対応を行うことを専門とする組織。
OT	Operational Technology の略称。運用技術と訳される。主に装置や工程の制御や監視などの運用を行うための技術を指す。
DDoS 攻撃	Distributed Denial of Service attack の略称。攻撃対象のウェブサイトやサーバに対して複数のコンピューターから攻撃を実施しサービス提供を困難にさせるサイバー攻撃。
ゼロデイ攻撃	OS やソフトウェアの脆弱性に対する修正プログラムが提供される前にその脆弱性を利用して行われる攻撃。
NGAV	Next Generation Anti-Virus の略称。「次世代アンチウイルス」などと訳される。従来のアンチウイルスに加え、振る舞い検知、AI による検知などの技術が導入されたソリューション。
EDR	Endpoint Detection and Response の略称。アンチウイルスソフトなどで防御しきれず、侵入を許してしまったマルウェアなどが動作する際の振る舞いを検知し、場合によっては動作や通信を遮断するソリューション。

WAF	Web Application Firewall の略称。主にアプリケーション層の通信の不正や異常を検知・防御するソリューション。製品により IDS・IPS がカバーする範囲が重複している場合もある。
IDS・IPS	IDS : Intrusion Detection System の略称。侵入検知システムと呼ばれ、不正や異常な通信を検知し管理者に通知するソリューション。 IPS : Intrusion Prevention System の略称。侵入防御システムと呼ばれ、不正や異常な通信を検知し防御・遮断を実施するソリューション。 IPS は不正通信検知後、通信遮断を実施するところが IDS と異なる。
NGFW	Next Generation Fire Wall の略称。通常のファイアウォールは IP アドレスやポート番号などで通信状況を確認するが、NGFW はアプリケーションレベルの通信状況の確認が可能であり、従来のファイアウォールと比較してより細かく通信の制御をすることが可能となっている。
SIEM	ファイアウォールや IDS・IPS などから出力されるログやデータを一元的に集約し、それらのデータを組み合わせて相関分析を行うことができるソリューション。サイバー攻撃などの際に、単独の製品のログやデータを用いるよりも通信状況の流れが詳細に把握できる。
UEBA	User and Entity Behavior Analysis の略称。ログやトラフィックからユーザーや機器の振る舞いを監視し、異常を検知するソリューション。ログやトラフィックの収集のために SIEM と連携する場合や、SIEM のオプションとして提供される場合などがある。
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart の略称。「人間とマシンを判別するチューリングテスト」のことである。歪んだ文字や数字などの画像を用いることで人間には理解できるがプログラムでは解析しにくい情報を提示し、正しく答えられるか確認する機能。スパム目的などの機械アクセスを防止するために提供される。
AI 活用ペネトレーションテスト	対象システムに AI が外部からの侵入を試みることによって脆弱性を発見する。AI が情報収集や攻撃に用いられるリクエストを送り、レスポンスを分析することで脆弱性の有無、攻撃の成否を評価する。 人間による検査と併用される場合が多い。
ソースコード診断	プログラムのソースコードを AI で診断する。ソースコードを読み込んで脆弱性の有無を確認する、ソースコードを画像として読み込んで可読性を評価する、などのサービスがある。